École Centrale de Nantes **Université de Nantes** **École des Mines de Nantes**

ÉCOLE DOCTORALE S T I M

« SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DES MATÉRIAUX »

Année 2007 N° ED 366-321

THÈSE DE DOCTORAT

Spécialité : INFORMATIQUE

présentée et soutenue publiquement par

Jérôme DAVID

le 8 novembre 2007
à l'École Polytechnique de l'Université de Nantes

AROMA : une méthode pour la découverte d'alignements orientés entre ontologies à partir de règles d'association

Président :	Jérôme EUZENAT, Directeur de recherche à L'INRIA Rhône-Alpes
Rapporteurs :	Chantal REYNAUD, Professeur à l'IUT d'Orsay, université Paris Sud Djamel ZIGHED, Professeur à l'université Lumière Lyon 2
Examinateurs :	Henri BRIAND, Professeur à l'école polytechnique de l'université de Nantes Fabrice GUILLET, Maître de conférences HDR à l'école polytechnique de l'université de Nantes Pascale KUNTZ, Professeur à l'école polytechnique de l'université de Nantes
Directeur de thèse :	Henri BRIAND, Professeur à l'école polytechnique de l'université de Nantes
Co-encadrant :	Fabrice GUILLET, Maître de conférences HDR à l'école polytechnique de l'université de Nantes
Laboratoire :	Laboratoire d'Informatique de Nantes Atlantique (LINA) 2, rue de la Houssinière – BP 92208 – 44322 Nantes Cedex 3

Je tiens à remercier Henri Briand qui a dirigé mes travaux de thèse. Il a su me guider et me faire part de son expérience tout en me laissant une grande liberté dans mes choix. Je remercie mon encadrant Fabrice Guillet pour ses nombreux conseils et son soutien. Grâce à lui j'ai beaucoup appris sur la recherche et l'enseignement. Je lui en suis très reconnaissant. Je tiens à remercier particulièrement Régis Gras pour sa gentillesse, son aide, son soutien, et ses conseils. J'ai beaucoup apprécié ses qualités intellectuelles et humaines. Je remercie Pascale Kuntz de m'avoir accueilli au sein de son équipe. Je remercie Jacques Philippé qui m'a poussé vers la recherche et qui m'a permis d'obtenir une bourse de la *fondation VediorBis pour la Recherche et l'Emploi* pour le financement de ma thèse. Je remercie Mme Chantal Reynaud et Mr Djamel Zighed d'avoir accepté d'être les rapporteurs de ma thèse, et pour leurs évaluations rigoureuses de mon manuscrit. Je remercie Mr Jérôme Euzenat d'avoir accepté d'être membre de mon jury.

Je remercie Catherine Galais pour les relectures et corrections de mes articles anglophones. Je remercie également tous les membres et thésards de l'équipe COD et notamment Xavier Aimé, Emmanuel Blanchard, Nicolas Beaume, Stéphane et Hélène Daviet, Julien Lorec, Fabien Picarougne, Bruno Pinaud pour leur contribution à l'ambiance de travail sereine et stimulante dans laquelle j'ai effectué cette thèse. Je remercie particulièrement Julien Blanchard pour ses nombreux conseils, ses remarques pertinentes et son soutien dans tous les moments difficiles. Je t'en suis vraiment très reconnaissant. Je remercie l'équipe du département Informatique de Polytech'Nantes et notamment Corinne Lorentz et Sophie Lanier pour leur gentillesse et leur soutien. Enfin, je remercie ma famille et mes amis pour m'avoir encouragé dans mes choix et particulièrement mon père pour la relecture de mon manuscrit.

Table des matières

Liste des figures

Liste des tableaux

Introduction

Extraction de connaissances dans les données : le modèle des règles d'association

Depuis l'apparition, dans les années 1960, de l'informatique et des bases de données, les volumes d'information stockée n'ont cessé de croître[1].

Au début des années 1990, la quantité, déjà colossale, d'information stockée a motivé le développement d'un nouveau courant de recherche appelé extraction des connaissances dans les données. L'extraction des connaissances dans les données (ECD), ou fouille de données a été définie comme l'extraction non-triviale d'information explicite, précédemment inconnue, et potentiellement utile, à partir de grands ensembles de données [FPSM91]. Son objectif principal est donc de fournir des connaissances potentielles à un utilisateur, en principe expert du domaine, mais non spécialiste des techniques de fouilles de données. Dans ce sens, une attention toute particulière est portée envers l'intelligibilité des résultats produits.

Un processus d'ECD se déroule typiquement en trois étapes successives [FPSS96] :

1. Le prétraitement des données consistant à préparer les données au processus de fouille.
2. La fouille de données qui consiste à appliquer un (ou des) algorithme(s) d'ECD sur les données prétraitées dans le but d'en extraire des motifs ou modèles.
3. Le post-traitement des résultats de fouille. Cette étape regroupe l'évaluation et la présentation à l'utilisateur des résultats qui, s'ils sont validés, deviendront des connaissances.

Une des principales méthodes d'ECD est l'extraction des règles d'association, introduite par [AIS93]. Les règles d'association sont des propositions de la forme « Si *prémisse* alors *conclusion* », notées *prémisse* $\rightarrow$ *conclusion*, où la prémisse et la conclusion sont des propositions portant sur les attributs d'une table issue d'une base de données. Une des toutes premières applications du modèle des règles d'association fut l'étude du panier de la ménagère. Elle consiste à étudier

[1] Selon une étude parue début 2007 du cabinet IDC (International Data Group), 161 exaoctets (161 milliards de gigaoctets) d'informations numériques ont été créées et copiées en 2006 [IDC07].

les tendances d'achat des clients d'un supermarché. Un exemple de règle plausible est « Si un client achète des fruits de mer alors il a tendance à acheter également du vin blanc ».

Un des principaux avantages des règles d'association est qu'elles représentent les connaissances de manière simple et explicite. Cela a motivé de nombreuses recherches qui ont abouti à la publication de nombreux algorithmes d'extraction de règles (voir [CR06] pour un état de l'art récent) dont le premier fut *Apriori* [AS94]. Cependant, une de leurs principales limites concerne les quantités prohibitives de règles générées par les algorithmes d'extraction.

Il est ainsi très difficile pour l'utilisateur de rechercher et de sélectionner parmi ces grands ensembles de règles, celles qui pourraient l'intéresser. De ce fait, des solutions visant à aider l'utilisateur dans cette démarche de post-traitement ont été envisagées. Une des solutions réside dans l'utilisation de mesures de qualités (ou d'intérêt) [Bla05], [BSGG04], [Fre98], qui permettent, selon un certain point de vue, de quantifier la qualité des règles et de les ordonner. Une seconde solution consiste à identifier et éliminer certaines formes de redondance dans les ensembles de règles extraites [PTB+05], [HF99], [Leh00], [SA95].

De l'ingénierie des connaissances à l'alignement d'ontologies

L'ingénierie des connaissances (IC) désigne un ensemble de concepts et techniques visant à acquérir, formaliser (et structurer), opérationnaliser les informations et connaissances d'un domaine dans le but de les manipuler et les diffuser entre les différents acteurs (humains ou artificiels) d'une organisation ou d'une communauté. Parmi les nombreuses disciplines impliquées dans l'IC, la représentation des connaissances [BL04], domaine de l'Intelligence Artificielle (IA), s'intéresse à la manière de représenter symboliquement les connaissances et leur sémantique dans le but de faciliter leur manipulation et leur réutilisation par des systèmes à base de connaissances. Depuis une quinzaine d'années, une attention particulière a été portée sur le partage et la réutilisation des représentations de connaissances. Ces efforts ont permis le développement des ontologies [Gru93], dénotant, au sens large, des représentations partagées et communes des connaissances d'un domaine qui peuvent être diffusées entre les humains et les programmes informatiques.

Le terme ontologie (en informatique) a reçu de nombreuses définitions. L'une des plus citée dans la littérature est celle de Gruber [Gru93] : « Une ontologie est une spécification explicite (et donc formelle) d'une conceptualisation ». Ainsi, une ontologie peut être vue comme une représentation abstraite des connaissances d'un domaine. Cette représentation étant formelle, elle doit être écrite dans un langage ou formalisme présentant une syntaxe et une sémantique précise (contrairement au langage naturel). Dans la pratique, une ontologie est un schéma représentant les connaissances d'un domaine au travers d'ensembles structurés de concepts et de propriétés. Cette structuration vient d'une part de la définition des propriétés mettant en relation les concepts eux-mêmes et d'autre part, de la relation de spécialisation/généralisation entre propriétés et

entre concepts.

Aujourd'hui, au sein du Web, ces modèles et technologies issus de l'IC sont devenus essentiels pour faire face aux volumes croissants d'information disponible, et donc aux besoins de solutions de structuration, de recherche et d'échange d'information et de connaissances. A la fin des années 1990, les efforts du W3C se sont portés sur le méta-langage XML [W3C88] qui joue, aujourd'hui, un rôle majeur dans la publication électronique de masse et dans l'échange d'une large variété de données sur le Web et partout ailleurs. Depuis plus récemment, le W3C promeut la vision du Web Sémantique [BHL01] dans laquelle l'information est donnée de manière explicite facilitant ainsi son traitement informatisé et son intégration sur le Web. Cette technologie sémantique est centrée autour de deux briques principales que sont RDF [W3C04b] et OWL [W3C04a]. RDF est utilisé pour représenter l'information et pour échanger des connaissances sur le Web. OWL est, quant à lui, utilisé pour publier et partager des ontologies ayant pour buts de faciliter la recherche avancée sur le Web, la communication entre agents logiciels et la gestion des connaissances.

Même si ces modèles issus de l'IC, et notamment les ontologies, facilitent l'organisation et donc l'échange et le partage des données et connaissances, le Web demeure un environnement hétérogène et dispersé. Cette hétérogénéité est due à la présence de nombreux formalismes de représentation (des simples taxonomies, aux ontologies OWL, en passant par des thésaurus SKOS) et la présence simultanée de taxonomies, ou ontologies, portant sur les mêmes domaines. Dans le but de résoudre, en partie, ces problèmes d'interopérabilité et d'intégration, une des solutions réside dans la comparaison des représentations de données et de connaissances. Dans ce sens, l'alignement de hiérarchies (i.e. de taxonomies, thésaurus, ou encore d'ontologies) vise à identifier les correspondances entre les entités (catégories, concepts, classes, propriétés) issues de deux représentations structurées, en mettant en exergue, notamment, la relation sémantique qu'elles entretiennent. Cette relation sémantique peut être, premièrement, du type équivalence (la catégorie « Humain » est équivalente à la catégorie « Homo Sapiens »), mais également du type implication (la catégorie « Enfant » est incluse dans la catégorie « Humain »).

Depuis le début des années 2000, on a pu remarquer un engouement certain envers les méthodes d'alignement. On recense, aujourd'hui, près d'une cinquantaine de systèmes [ES07] issus de diverses communautés telles que les bases de données, la recherche d'information, le traitement du langage, l'ingénierie des connaissances, l'apprentissage, etc. Il ressort, globalement, que la grande majorité des méthodes d'alignement sont basées sur des combinaisons, plus ou moins complexes, de mesures de similarité. De ce fait, on remarque deux limites majeures :

- La sémantique des alignements produits est limitée à l'équivalence. En effet, de par leur nature symétrique, les mesures de similarités ne peuvent pas être utilisées pour détecter des relations d'implication (ou subsomption).
- Les combinaisons complexes de mesures, même si elles permettent d'obtenir de meilleures performances, rendent les résultats difficilement interprétables pour l'utilisateur. En effet, ce dernier est en droit de se demander pourquoi deux entités sont en correspondance dans l'alignement.

Un autre constat peut également être fait concernant le peu d'efforts qui ont été consacrés pour l'aide à la visualisation, à la validation, et à l'édition des alignements.

Contributions de la thèse

Afin de dépasser les limites dues, notamment à la sémantique réduite des alignements, et à la difficulté d'interprétation des résultats, cette thèse a pour principale ambition de tirer profit des travaux menés sur la fouille de règles d'association. Plus particulièrement, en nous appuyant sur ce modèle simple, intelligible et muni de mesures de qualité, nous proposons d'enrichir la sémantique des alignements en permettant la détection et l'évaluation de quasi-implications. Les contributions de cette thèse se déclinent en 5 points dont un majeur concernant la proposition d'une méthode originale d'alignement extensionnelle, terminologique, et orientée.

1. Un modèle d'alignement implicatif
Nous présentons les modèles de hiérarchies et d'alignement en prenant soin de traiter les aspects engendrés par la prise en compte d'implication, à savoir, la redondance, la cardinalité et la symétrie des alignements.

2. Une méthode d'alignement
En nous appuyant sur notre modélisation de l'alignement, nous proposons la méthode d'alignement *AROMA* (Association Rule Ontology Matching Approach). L'idée sous-jacente à notre approche est que deux entités x et y sont en relation d'implication, $x \Rightarrow y$, si le vocabulaire utilisé dans les descriptions et les instances de x, a tendance à être inclus dans celui de y. Cette méthode est conçue suivant le schéma classique, en trois étapes, des approches d'ECD :

1. **Pré-traitement des hiérarchies :** Nous proposons deux méthodes de pré-traitement permettant de repeupler les hiérarchies à comparer sur un ensemble de vocabulaire commun. La première est dédiée aux hiérarchies textuelles, la seconde est adaptée aux ontologies décrites en RDFS/OWL. Ces deux méthodes de pré-traitement procèdent à l'acquisition des termes contenus dans les descriptions textuelles en s'appuyant sur des outils de TAL (Traitement Automatique du Langage). La première méthode permet, en outre, de sélectionner, pour chaque entité, un ensemble de termes dits représentatifs.
2. **L'extraction d'alignement implicatif :** Cette deuxièmes phase permet, à partir des hiérarchies redéfinies sur des termes, d'évaluer et d'extraire les règles d'association entre entités. Ces règles nous permettent de produire un alignement qui est orienté, car prenant en compte des relations de correspondance de type implication.
3. **Post-traitements d'alignement :** Dans l'objectif de produire des alignements consistants et minimaux (non-redondants), nous proposons une série de filtres permettant (1) de détecter les inconsistances et de les éliminer ; (2) de supprimer les redondances ; (3) d'adapter la cardinalité et la symétrie de l'alignement produit.

Nous proposons également une méthode intensionnelle d'alignement syntaxique qui permet, a posteriori, d'enrichir l'alignement avec des correspondances non détectées par la méthode extensionnelle.

3. Un modèle d'évaluation adapté aux alignements orientés
L'évaluation des performances des méthodes d'alignement est couramment basée sur le modèle classique d'évaluation utilisé en recherche d'information (mesures de précision et de rappel) [vR79]. Nous montrons que ce modèle est inadapté à l'évaluation des alignements. A partir du modèle d'évaluation sémantique de J. Euzenat [Euz07], nous proposons une adaptation de ce dernier modèle pour une meilleure prise en compte de la relation d'implication dans un alignement.

4. Réalisation d'un système d'aide à l'alignement
Poursuivant notre idée de proposer une démarche d'ECD dédiée à l'alignement, nous proposons d'intégrer notre méthode d'alignement dans un système interactif d'aide à l'alignement. Plus particulièrement, nous proposons une interface de visualisation, d'aide à la validation et à l'édition d'alignements.

5. Expérimentations
Nous nous sommes également attachés à l'étude expérimentale du comportement et de la performance de notre méthode AROMA sur différents types de hiérarchies et avec différentes mesures de qualité. En effet, nous avons réalisé nos tests tant sur des hiérarchies textuelles que sur des ontologies décrites en RDFS/OWL. Avec chacun des jeux de tests, nous avons utilisé 6 mesures d'intérêt différentes et étudié les résultats obtenus.

Organisation de la thèse

Cette thèse est organisée en 6 chapitres. Dans le chapitre 1, nous présentons le modèle des règles association en insistant, d'une part, sur les mesures d'intérêt et d'autre part, sur les techniques de réduction de redondance appliquées à la fouille de règles entre attributs organisés en taxonomies.

Le chapitre 2 introduit les modèles de hiérarchie et d'alignement à partir desquels nous formalisons la suite de notre travail. Ce chapitre met notamment en avant les notions de déduction, de redondance, de consistance et de symétrie d'un alignement. Finalement, nous proposons une formalisation du contexte de fouille de règles adaptée à l'alignement de hiérarchies.

Le chapitre 3 présente une étude structurée des techniques d'alignement. Tout d'abord, nous nous concentrons sur les caractéristiques externes (entrées/sorties) et la composition interne des méthodes d'alignement. Ensuite, nous focalisons sur les techniques utilisées aux sein des algorithmes d'alignements, en distinguant celles qui sont utilisées sur la description intensionnelle d'une hiérarchie de celles qui s'appuient sur l'extension.

Dans le chapitre 4, nous présentons la méthode d'alignement AROMA, en détaillant successivement les trois phases de l'approche : prétraitements, extraction de règles, et post-traitements.

Dans le chapitre 5, nous fournissons une présentation du modèle classique

d'évaluation des méthodes d'alignement et une critique de ce dernier faisant ressortir ses limites tant sur les aspects de prise en compte de la sémantique des hiérarchies et des alignements. Ensuite, à partir du modèle sémantique proposé par [Euz07] et des notions de déduction et de fermeture d'alignement présentées dans le chapitre 2, nous proposons une adaptation du modèle d'évaluation prenant mieux en compte le caractère potentiellement orienté d'un alignement.

Finalement, dans le chapitre 6, nous présentons, tout d'abord, la réalisation logicielle d'AROMA ainsi que de son extension AROMAViz qui est une interface d'aide à la visualisation, à la validation, et à l'édition d'alignement. Ensuite, nous proposons des évaluations expérimentales d'AROMA. Ces expérimentations sont découpées en trois volets : (1) Evaluation du pré-traitement des hiérarchies textuelles, plus particulièrement de la phase de sélection des termes ; (2) Evaluation d'AROMA pour l'alignement de hiérarchies textuelles ; (3) Evaluation d'AROMA pour l'alignement d'ontologies RDFS/OWL.

Règles d'association et mesures d'intérêt

1

Sommaire

Introduction

En ECD, la fouille de règles d'association [AIS93] est une technique populaire produisant des connaissances de la forme « Si *prémisse* alors *conclusion* », notées *prémisse* $\rightarrow$ *conclusion*. Dans une règle d'association, la prémisse et la conclusion sont des propositions portant sur les attributs d'une table. Une règle *prémisse* $\rightarrow$ *conclusion* représente une tendance implicative entre l'ensemble des enregistrements vérifiant la prémisse vers l'ensemble des enregistrements vérifiant la conclusion.

Un des principaux atouts des règles d'association est qu'elles représentent des connaissances de manière simple et explicite. De par la nature non-supervisée des algorithmes d'extraction de règles d'association, cette technique d'apprentissage n'a pas besoin qu'on lui fournisse d'information particulière sur les connaissances à découvrir, contrairement aux techniques supervisées telles que les arbres décision [ZR00] ou les réseaux bayésiens [NWL$^+$04]. Cependant, cette nature non-supervisée constitue également l'une des principales limites de la méthode.

En effet, la quantité de règles générées par les algorithmes ([AIS93], [AS94]) croît de manière exponentielle en fonction du nombre d'attributs considérés. Afin de faire face aux quantités impressionnantes de règles générées, des méthodes permettant de filtrer les règles ont été proposées.

Une première famille de méthodes est constituée des mesures d'intérêts. Ces mesures permettent non seulement de vérifier la qualité implicative des règles, mais également d'étudier de nombreuses caractéristiques telles que la nouveauté, la significativité, la surprise, la non-trivialité, l'actionabilité [CR06]. Les mesures d'intérêt permettent ainsi de classer et de filtrer les règles produites, et par conséquent, elles aident l'utilisateur à choisir les meilleures règles en fonction de ses préférences.

Une autre famille de méthodes concerne la réduction des redondances dans l'ensemble des règles extraites. L'élimination des redondances a été étudiée, tout d'abord, par l'introduction des règles d'association généralisées [SA95], [SA97], [HF99]. Les règles d'association généralisées sont une extension du modèle classique qui prend en compte une taxonomie d'items (ou hiérarchie de concepts). Ce nouveau modèle a été étudié au niveau des algorithmes d'extraction. D'autres méthodes ont été proposées dans le cadre du post-traitement par [Leh00], [PTB+05]. Ces méthodes sont basées sur l'approche des dépendances fonctionnelles (axiomes d'Amstrong, fermeture transitive, couverture minimale).

Après une présentation du modèle des règles d'association, ce chapitre s'intéresse à une étude des propriétés et une classification des mesures d'intérêt. Ensuite, le modèle des règles d'association généralisées et l'approche de réduction des redondances seront étudiés selon notre objectif de fouille de règles entre hiérarchies de concepts.

1.1 Définitions et notations

Nous considérons un ensemble E composé de n individus décrits par un ensemble I de variables booléennes au moyen de la relation binaire $\delta \subseteq I \times E$. L'ensemble des individus associés à une variable booléenne $i \in I$ est noté $\delta(i)$.

Exemple. E peut représenter l'ensemble des paniers vendus d'un supermarché (c.-à-d. l'ensemble des tickets de caisse) et $I = \{$*huîtres, muscadet...*$\}$ l'ensemble des produits vendus dans ce supermarché. L'ensemble des paniers contenant du muscadet est noté $\delta(\mathit{muscadet})$, l'ensemble des paniers ne contenant pas de muscadet est noté $\delta(\overline{\mathit{muscadet}})$.

Par extension, si l'on considère un ensemble $a \subseteq I$ de variables booléennes, appelé itemset, l'ensemble des individus $x \in E$ associés par δ à l'ensemble des variables de a est noté $A = \bigcap_{i \in a} \delta(i)$. La cardinalité de A est notée n_a.

Exemple. Si $a = \left\{\mathit{muscadet}, \overline{\mathit{huître}}\right\}$, l'ensemble A désigne l'ensemble des paniers contenant du muscadet mais pas d'huître.

Définition 1.1 *Une règle d'association est un couple de variables noté $a \rightarrow b$ où a et b sont des itemsets disjoints, appelés respectivement prémisse et conclusion.*

Une règle d'association représente une tendance implicative entre l'ensemble

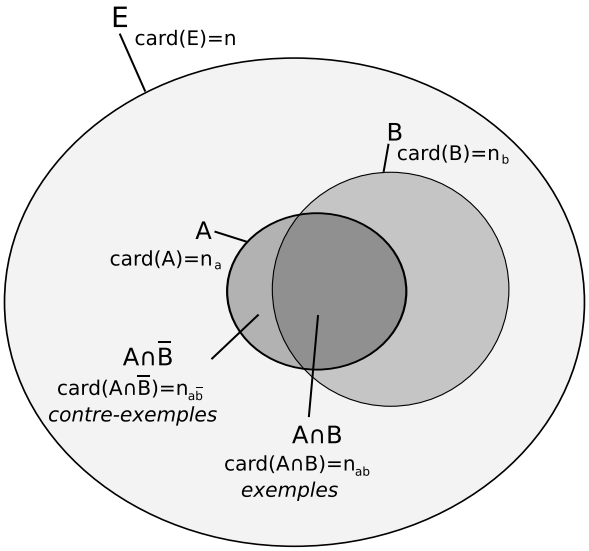


FIG. 1.1 – Diagramme de Venn d'une règle $a \rightarrow b$

des individus associés à a vers l'ensemble des individus associés à b. Elle peut se lire de la manière suivante [Bla05] : « Si un individu vérifie a alors il vérifie sûrement b ». Par exemple, une règle $a \rightarrow b$ sur l'ensemble de paniers vendus par un supermarché signifiera : « Si un panier contient l'ensemble des produits a alors il aura tendance à contenir l'ensemble des produits b ». Les exemples d'une règle, notés $a \cap b$, représentent les individus associés à la fois à a et b. Les contre-exemples de la règle, notés $a \cap \overline{b}$, sont les individus associés à a mais pas à b. Une règle est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples. Une règle n'ayant que des exemples est une implication logique.

Pour une règle $a \rightarrow b$, on retient 3 règles qui lui sont liées :
- sa contraposée : $\overline{b} \rightarrow \overline{a}$,
- sa réciproque : $b \rightarrow a$,
- son contraire : $\overline{a} \rightarrow b$.

1.2 Mesures d'intérêt

1.2.1 Support et Confiance

Deux mesures sont utilisées usuellement lors de la fouille de règles d'association, il s'agit du support et de la confiance. Le support représente la fréquence d'individus respectant la règle dans l'ensemble étudié [AIS93] :

$$support(a \rightarrow b) = \frac{n_{ab}}{n}$$

Exemple. Reprenons la règle *huître* $\rightarrow$ *muscadet*. Si son support est égal

à $0,8$ ($support(huître \rightarrow muscadet) = 0,8$), cela signifie que 80% des paniers contiennent à la fois des huîtres et du muscadet.

La confiance permet quant à elle de vérifier la validité de la règle. Elle représente la proportion d'individus associés à la prémisse qui sont également associés à la conclusion [AIS93] :

$$confiance(a \rightarrow b) = \frac{n_{ab}}{n_a}$$

Exemple. Si la confiance de la règle *huître* $\rightarrow$ *muscadet* est égale à $0,8$, cela signifie que 80% des paniers contenant des huîtres, contiennent également du muscadet.

1.2.2 Caractéristiques et classification des mesures d'intérêt

Dans sa thèse [Bla05], J. Blanchard propose une classification des mesures d'intérêt. Cette classification, basée sur les propriétés des mesures, permet de mieux appréhender leur sémantique. Nous présentons dans cette section, cette classification qui distingue l'objet, la portée et la nature d'une mesure d'intérêt.

Objet d'une mesure

Entre les deux situations extrêmes $n_{ab} = min(n_a, n_b)$ et $n_{ab} = max(0, n_a + n_b - n)$, où une règle $a \rightarrow b$ sera respectivement de qualité maximale ou minimale, il existe deux situations particulières où cette règle peut être considérée, selon un certain point de vue, non orientée et donc pas intéressante. Ces deux situations particulières sont l'équilibre et l'indépendance.

Une règle est en situation d'équilibre lorsque elle a autant d'exemples que de contre-exemples. Dans ce cas, les nombres d'exemples et de contre-exemples seront égaux à $\frac{n_a}{2}$. Une règle est en situation d'indépendance lorsque les variables a et b sont indépendantes. Dans ce cas, la probabilité d'avoir simultanément la réalisation de a et b est $P(a \cap b) = P(a) \times P(b)$. Dans cette situation, le nombre d'exemples de la règle sera $n_{ab} = \frac{n_a . n_b}{n}$.

Une mesure d'intérêt va quantifier soit un écart à l'équilibre, soit un écart à l'indépendance.

Définition 1.2 *Une mesure d'intérêt est une mesure d'écart à l'équilibre, et notée m_e, si elle prend une valeur fixe, notée v_e, à l'équilibre.*

$$m_e(a \rightarrow b) = v_e \equiv n_{ab} = \frac{n_a}{2}$$

Un exemple de mesure d'écart à l'équilibre est la confiance. En effet, lorsque $n_{ab} = \frac{na}{2}$, la confiance prend une valeur fixe $v_e = 0,5$.

Définition 1.3 *Une mesure d'intérêt est une mesure d'écart à l'indépendance, et notée m_i, si elle prend une valeur fixe, notée v_i, à l'indépendance.*

$$m_i(a \rightarrow b) = v_i \equiv n_{ab} = \frac{n_a . n_b}{n}$$

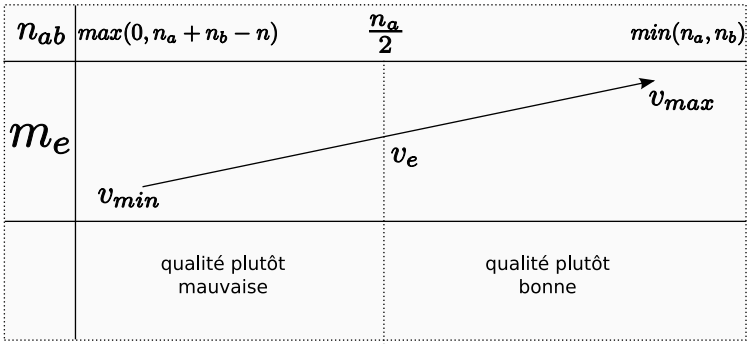


FIG. 1.2 – Variation d'une mesure d'écart à l'équilibre

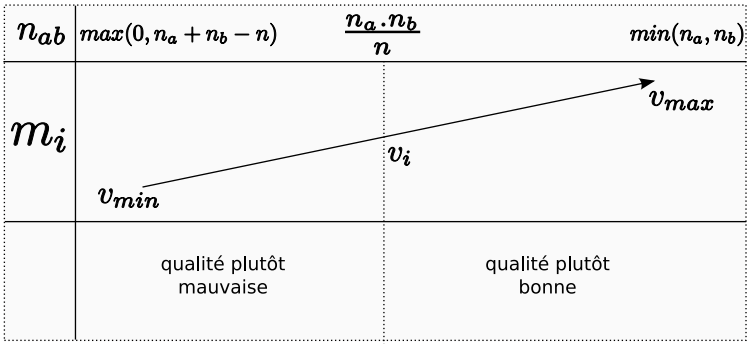


FIG. 1.3 – Variation d'une mesure d'écart à l'indépendance

Un exemple de mesure d'écart à l'indépendance est le lift défini ainsi :

$$lift(a \rightarrow b) = \frac{n.n_{ab}}{n_a.n_b}$$

Le lift prend une valeur fixe $v_i = 1$ à l'indépendance.

La valeur d'une mesure d'intérêt sera d'autant plus élevée que le nombre d'exemples s'éloigne de la situation dont elle mesure l'écart (équilibre ou indépendance) et se rapproche de la situation de qualité maximale $n_{ab} = min(n_a, n_b)$. De manière duale, la valeur d'une mesure d'intérêt sera d'autant plus faible que le nombre d'exemples s'éloigne de la situation dont elle mesure l'écart et se rapproche de la situation de qualité minimale $n_{ab} = max(0, n_a + n_b - n)$. Les variations des mesures d'écart à l'indépendance et d'écart à l'équilibre sont schématisées sur les figures 1.2 et 1.3.

Une mesure d'écart à l'équilibre et une mesure d'écart à l'indépendance étudient deux aspects différents (mais complémentaires) d'une règle. En effet, si une règle $a \rightarrow b$ est bonne en terme d'écart à l'équilibre, cela signifie : « si a est vérifié alors b a de grandes chances de l'être également ». Une règle bonne en terme d'écart à l'indépendance signifie : « si a est vérifié alors b a plus de chances de l'être également (qu'à l'accoutumée) ».

Exemple. Reprenons l'exemple des paniers vendus par un supermarché et la

règle *huître* $\rightarrow$ *muscadet*. Si cette règle a une confiance égale à $0,8$ alors elle sera bonne en terme d'écart à l'équilibre. Cependant, si l'on remarque que 80% des paniers contiennent de toute façon du muscadet alors cette règle est en situation d'indépendance ($lift($*huître* $\rightarrow$ *muscadet*$) = 1$) : le fait qu'un panier contienne des huîtres n'augmente pas les chances qu'il contienne du muscadet. Cette règle est donc mauvaise en terme d'écart à l'indépendance.

Portée d'une mesure

Une mesure d'intérêt peut également mesurer une ressemblance entre la règle $a \rightarrow b$ et une configuration logique telle que l'implication $a \Rightarrow b$, la conjonction $a \wedge b$ ou l'équivalence $a \Leftrightarrow b$.

Nous avons vu que, pour une règle $a \rightarrow b$, ses exemples sont les individus vérifiant à la fois a et b, notés $A \cap B$, et que ses contre-exemples sont les individus qui vérifient a mais pas b, notés $A \cap \overline{B}$. Cependant, les individus qui vérifient b mais pas a, notés $\overline{A} \cap B$, et les individus qui ne vérifient ni a, ni b, notés $\overline{A} \cap \overline{B}$ ne sont, dans ce cas, pas considérés. Si une mesure associe une sémantique d'exemple ou de contre-exemple à certains de ces ensembles d'individus alors elle se focalise sur un cas particulier de règle qui peut être la quasi-implication, la quasi-conjonction ou la quasi-équivalence.

Une quasi-implication est une règle, notée $a \Rightarrow b$, pour laquelle les individus $\overline{A} \cap \overline{B}$ sont considérés comme des exemples. Une quasi-implication $a \Rightarrow b$ est ainsi équivalente à sa contraposée $\overline{b} \Rightarrow \overline{a}$. A partir de cette équivalence, on définit une mesure de quasi-implication de la manière suivante :

Définition 1.4 *Une mesure de quasi-implication est une mesure m vérifiant* $m(a \rightarrow b) = m(\overline{b} \rightarrow \overline{a})$.

Une quasi-conjonction est une règle, notée $a \wedge b$, pour laquelle les individus $A \cap \overline{B}$ constituent également des contre-exemples. Une quasi-conjontion $a \wedge b$ est ainsi équivalente à sa réciproque $b \wedge a$. Une mesure de quasi-conjonction est définie ainsi :

Définition 1.5 *Une mesure de quasi-conjonction est une mesure m vérifiant* $m(a \rightarrow b) = m(b \rightarrow a)$

Une quasi-équivalence est une règle, notée $a \Leftrightarrow b$, pour laquelle les individus $\overline{A} \cap \overline{B}$ sont considérés comme des exemples, et les individus $A \cap \overline{B}$ constituent des contre-exemples. Une quasi-équivalence est à la fois une quasi-implication et une quasi-conjonction. Elle est ainsi équivalente à sa contraposée et à sa réciproque.

Définition 1.6 *Une mesure de quasi-équivalence est une mesure m vérifiant* $m(a \rightarrow b) = m(b \rightarrow a) = m(\overline{b} \rightarrow \overline{a}) = m(\overline{a} \rightarrow \overline{b})$

Ce critère de classification des mesures d'intérêt selon leur portée est synthétisé par la table 1.1.

Portée	**Notation**	**Exemples**	**Contre-exemples**	**Règles équivalentes**
Règle	$a \rightarrow b$	$A \cap B$	$A \cap \overline{B}$	
Quasi-implication	$a \Rightarrow b$	$A \cap B$, $\overline{A} \cap \overline{B}$	$A \cap \overline{B}$	$\overline{b} \rightarrow \overline{a}$
Quasi-conjonction	$a \wedge b$	$A \cap B$	$A \cap \overline{B}$, $\overline{A} \cap B$	$b \rightarrow a$
Quasi-équivalence	$a \Leftrightarrow b$	$A \cap B$, $\overline{A} \cap \overline{B}$	$A \cap \overline{B}$, $\overline{A} \cap B$	$b \rightarrow a$, $\overline{b} \rightarrow \overline{a}$, $\overline{a} \rightarrow \overline{b}$

TAB. 1.1 – Portée d'une mesure d'intérêt

Nature d'une mesure

Un dernier critère de distinction des mesures d'intérêt concerne leur nature qui peut être descriptive ou statistique.

Une mesure est de nature descriptive si elle ne varie pas avec la dilatation des effectifs. En d'autres termes, si l'on considère une mesure m comme fonction de quatre paramètres n_a, n_b, n_{ab} et n, alors une mesure m descriptive vérifie $m(n_a, n_b, n_{ab}, n) = m(\alpha.n_a, \alpha.n_b, \alpha.n_{ab}, \alpha.n)$. Avec une mesure descriptive, les cardinaux sont considérés de manière relative.

Une mesure est de nature statistique si elle varie avec la dilatation des effectifs. Une mesure statistique m ne vérifie pas l'égalité : $m(n_a, n_b, n_{ab}, n) = m(\alpha.n_a, \alpha.n_b, \alpha.n_{ab}, \alpha.n)$. Ainsi, une mesure statistique va considérer les cardinaux de manière absolue. Avec ce type de mesure, une règle sera d'autant de meilleure qualité qu'elle est observée sur un grand volume de données.

Classification

La table 1.2 présente la classification, proposée par [Bla05], des principales mesures d'intérêt. Dans cette classification, on peut distinguer une ligne regroupant des mesures de similarité[1]. De part leur nature symétrique, ces mesures ne peuvent pas avoir une portée qui soit la règle, au sens strict, ou la quasi-implication. De plus, il n'existe pas de mesure de quasi-conjonction ou de quasi-équivalence qui a pour objet l'écart à l'équilibre.

[1] la définition de la similarité utilisée dans cette classification est celle de Lerman [Ler81] : une mesure de similarité est fonction $m_s(n_a b, n_{a\overline{b}}, n_{\overline{a}b}, n)$ de $\mathbb{N}^4$ dans $\mathbb{R}^+$ qui est symétrique en $n_{a\overline{b}}$ et $n_{\overline{a}b}$, croissante avec n_{ab} et décroissante avec $n_{a\overline{b}}$ lorsque que les autres variables sont fixes. Les variations sont strictes.

Portée / **Objet**	**Règle**	**Quasi-implication**	**Quasi-conjonction**	**Quasi-équivalence**
Ecart à l'équilibre	– confiance [AIS93], – indice de Sebag et Schoenauer [SS88], – taux des exemples et contre-exemples [Gui04], – estimateur laplacien de probabilité conditionnelle [BA99], – indice de Ganascia [Gan91], – moindre-contradiction [Aze03], – *indice probabiliste d'écart à l'équilibre (Ipee) [Bla05]*	– indice d'inclusion [GCB+04]		
Ecart à l'indépendance	– multiplicateur de cotes [LT04]	– indice de Loevinger [Loe47], – conviction [BMUT97] – *intensité d'implication* [Gra96], – *indice d'implication* [Gra96]	– lift ou intérêt [BMS97] – *vraisemblance du lien* [Ler81], – *contribution orientée au* χ^2 [Ler81]	– coefficient de corrélation [Pea96], – nouveauté [LFZ99], – collective strength [AY01], – κ [Coh60], – indice de Yule [Yul00], – rapport de cotes [Mos68] – *rule-interest* [PS91]
Similarité			– support ou indice de Russel et Rao [RR40], – indice de Jaccard [Jac01], – indice de Dice [Dic45], – indice d'Ochiai [Och57], – indice de Kulczynski [Kul27]	– support causal ou indice de Sokal et Michener [SM58], – indice de Rogers et Tanimoto [RT60]

La **nature** des mesures est indiquée par le style de la police : les mesures en *italique* sont statistiques, les autres sont descriptives.

Tab. 1.2 – Classification des mesures d'intérêt de [Bla05]

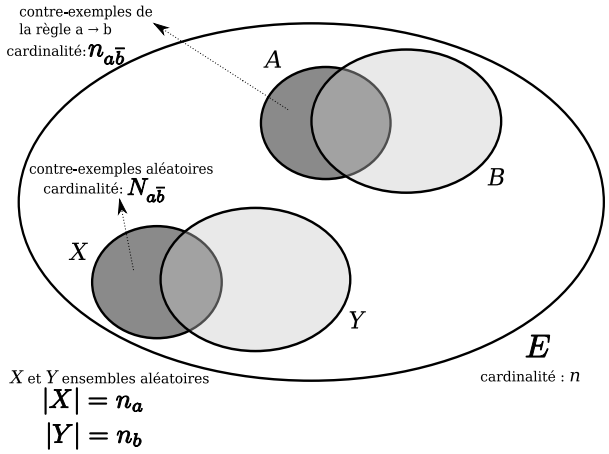


FIG. 1.4 – Règle $a \rightarrow b$ et tirage de deux ensembles X et Y indépendants

1.2.3 Intensité d'implication

L'intensité d'implication [Gra96] est un indice statistique d'écart à l'indépendance. Basée sur un modèle probabiliste, elle permet de comparer une distribution observée dans les données par rapport à une distribution théorique.

Modélisation

A l'instar de I.C. Lerman [Ler81], R. Gras propose, pour une règle $a \rightarrow b$, de comparer le nombre de contre-exemples observé, $n_{a\bar{b}}$ à celui attendu sous hypothèse H_0 d'indépendance entre a et b.

Pour cela, la formalisation consiste à considérer deux parties X et $\overline{Y}$ de E choisies, sous hypothèse H_0, aléatoirement et indépendemment l'une de l'autre et qui sont respectivement de même cardinalité que les ensembles A et $\overline{B}$ (illustrés figure 1.4). Soit la variable aléatoire $N_{a\bar{b}} = |X \cap \overline{Y}|$. L'intensité d'implication φ de la règle $a \rightarrow b$ représente la probabilité que le nombre $n_{a\bar{b}}$ soit plus petit que $N_{a\bar{b}}$:

$$\varphi(a \rightarrow b) = 1 - Pr\left[N_{a\bar{b}} \leq n_{a\bar{b}}\right]$$

La règle $a \rightarrow b$ sera dite admissible au seuil α si $\varphi(a \rightarrow b) \geq 1 - \alpha$

I.C. Lerman [Ler81] propose trois modélisations possibles pour exprimer H_0. Selon celle retenue la variable $N_{a\bar{b}}$ peut suivre :

- une loi hypergéométrique de paramètres $(n, n_a, n_{\bar{b}})$,
- une loi binomiale de paramètres $(n, \frac{n_a n_{\bar{b}}}{n^2})$,
- une loi de Poisson de paramètre $\lambda = \frac{n_a . n_{\bar{b}}}{n}$.

La loi hypergéométrique, de par ses symétries, ne permet pas de distinguer les règles $a \rightarrow b$ et $b \rightarrow a$. En effet, avec une telle modélisation, l'intensité d'implication est un indice de quasi-équivalence [Bla05]. La loi de Poisson est, quant à elle, celle qui maximise la dissymétrie entre les valeurs calculées pour

$a \rightarrow b$ et $b \rightarrow a$. De ce fait, nous adopterons, dans le cadre de cette thèse, la modélisation poissonienne.

Expression analytique

Ainsi, en retenant une modélisation de Poisson pour $N_{x\overline{y}}$ de paramètre $\lambda = \frac{n_a . n_{\overline{b}}}{n}$, la valeur de l'intensité d'implication est définie ainsi :

$$\varphi(a \rightarrow b) = 1 - e^{-\lambda} \sum_{k=0}^{n_{a\overline{b}}} \frac{\lambda^{-k}}{k!}$$

Lorsque $\lambda > 10$, la loi de Poisson peut être approximée par une loi normale. L'intensité d'implication s'écrit alors :

$$\varphi(a \rightarrow b) = \frac{1}{\sqrt{2\pi}} \int_{q(a,\overline{b})}^{\infty} e^{-\frac{t^2}{2}} dt$$

où $q(a, \overline{b}) = \frac{n_{a\overline{b}} - \lambda}{\sqrt{\lambda}}$. $q(a, \overline{b})$ est appelé indice d'implication [Gra96]. Cet indice quantifie la non-implication de a sur b.

Propriété

Lorsque n_b tend vers n (n_a, n_{ab}, et n fixés), alors $\varphi(a \rightarrow b)$ tend vers 0 [Gra96].

1.3 Fouille de règles généralisées et de règles non redondantes

Généralement les algorithmes d'extraction génèrent un ensemble prohibitif de règles dont beaucoup d'entre elles sont redondantes. Cette redondance peut être induite par la présence d'une relation de généralisation/spécialisation entre les prémisses (resp. conclusions) de règles a priori différentes. La relation de généralisation/spécialisation intervient à deux niveaux distincts. Premièrement, elle peut être présente sous forme d'une inclusion entre les ensembles de variables constituant les prémisses et conclusions. Par exemple, la conclusion de la règle "*huîtres* $\rightarrow$ *vin blanc*" sera plus générale que celle de la règle "*huîtres* $\rightarrow$ *vin blanc* $\wedge$ *citron*" puisque $\{$*vin blanc*$\} \subset \{$*vin blanc*, *citron*$\}$. Deuxièmement, cette relation de généralisation/spécialisation peut être définie sur les variables elles-mêmes. Par exemple, la prémisse de la règle "fruits de mer $\rightarrow$ vin blanc" est plus générale que celle de la règle "huîtres $\rightarrow$ vin blanc", puisque la variable *fruits de mer* est, par essence, plus générale que la variable *huîtres*.

Définition 1.7 *Une variable $i \in I$ est plus générale qu'une variable $j \in I$ sur E, notée $j \leq i$, si et seulement si $\delta(j) \subseteq \delta(i)$. Par extension, un itemset $a \subseteq I$ est plus général qu'un itemset $b \subseteq I$ sur E, noté $b \leq a$, si et seulement si $B \subseteq A$.*

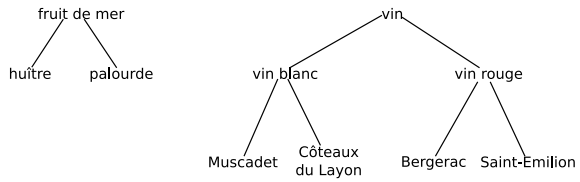


FIG. 1.5 – Exemple de taxonomies sur des variables d'une base de données

Une des propriétés découlant de cette notion de généralité entre attributs est que si un attribut (ou itemset) b est plus général qu'un attribut (ou itemset) a (noté $a \leq b$) alors la règle $a \rightarrow b$ est une implication logique et a donc une confiance $confiance(a \rightarrow b) = 1$.

A partir de ces observations, plusieurs approches de réduction du nombre de règles redondantes exploitant la relation de généralisation/spécialisation entre itemsets ont été proposées. La première tendance consiste à s'appuyer sur des taxonomies de variables [SA95], [HF99]. La deuxième tendance vise à supprimer les redondances en s'appuyant seulement sur l'inclusion entre les ensembles de variables [Leh00], [PTB+05].

1.3.1 Règles d'association généralisées

La notion de règles d'association généralisées a été introduite dans [SA95]. Les auteurs proposent dans ce papier d'extraire les règles intéressantes les plus générales en termes de support.

Définition 1.8 *Une règle $a \rightarrow b$ sera plus générale en terme de support qu'une règle $a' \rightarrow b'$ si $a' \leq a$ et $b' \leq b$.*

Etant donnée la présence possible d'une relation d'ordre sur l'ensemble des variables du jeu de données, la notion de règle d'association est généralisée pour prendre en compte la relation d'ordre.

Définition 1.9 *Une règle d'association généralisée $a \rightarrow b$ est une règle d'association pour laquelle il n'existe pas de variable de la conclusion qui soit plus générale qu'une variable de la prémisse. Formellement, $a \rightarrow b$ sera une règle d'association généralisée si l'assertion $\forall a_i \in a,\ \nexists b_j \in b,\ a_i \leq b_j$ est vérifiée [SA95].*

Exemple. La règle *huître* $\wedge$ *muscadet* $\rightarrow$ *vin blanc* n'est pas une règle généralisée puisque muscadet est plus spécifique que vin blanc. Son interprétation « si l'on achète des huîtres et du muscadet, alors le muscadet est du vin blanc » est complétement triviale et ainsi, cette règle est inutile. Par contre, la règle *huître* $\wedge$ *vin blanc* $\rightarrow$ *muscadet* sera une règle généralisée. Elle peut être interprétée « si l'on achète des huîtres et du vin blanc alors ce vin blanc sera sûrement du muscadet ».

La notion d'ancêtre d'un itemset z est définie comme étant un itemset plus général que z ayant le même nombre de variables.

Définition 1.10 *Un ensemble d'attributs $\widehat{z}$ est un ancêtre de l'ensemble d'attributs z si les contraintes suivantes sont vérifiées :*
- $|z| = |\widehat{z}|$ *: les deux ensembles z et $\widehat{z}$ contiennent le même nombre de variables.*
- $\widehat{z} \neq z$ *: les deux ensembles z et $\widehat{z}$ sont différents.*
- $\forall z_i \in \widehat{z} - z,\ \exists z_j \in z,\ z_j \leq z_i$ *: chaque variable de $\widehat{z}$ possède une variable plus spécifique ou identique dans z.*

Exemple. Les itemsets $\{huître, vin\ blanc\}$ et $\{fruit\ de\ mer, vin\ blanc\}$ sont des ancêtres de l'itemset $\{huître, muscadet\}$. Cependant, l'itemset $\{vin\ blanc\}$, même s'il est plus général que $\{huître, muscadet\}$, n'est pas un de ses ancêtres, puisqu'ils ne contiennent pas les mêmes nombres d'éléments.

A partir de la définition de la notion d'ancêtre d'un itemset, la notion d'ancêtre d'une règle d'association généralisée peut être définie.

Définition 1.11 *Une règle $c \rightarrow d$ est un ancêtre de la règle $a \rightarrow b$ si c est un ancêtre de a ou d est un ancêtre de b. Une règle $\widehat{a} \rightarrow \widehat{b}$ est l'ancêtre proche de la règle $a \rightarrow b$ si il n'existe pas de règle $a' \rightarrow b'$ telle que $a' \rightarrow b'$ est un ancêtre de $a \rightarrow b$ et $\widehat{a} \rightarrow \widehat{b}$ est un ancêtre de $a' \rightarrow b'$.*

Exemple. La règle *fruit de mer* $\rightarrow$ *vin* est un ancêtre de la règle *huître* $\rightarrow$ *muscadet* (car *huître* $\leq$ *fruit de mer* et *muscadet* $\leq$ *vin*). Cependant, elle n'est pas son ancêtre proche car la règle *fruit de mer* $\rightarrow$ *vin blanc*, dont *fruit de mer* $\rightarrow$ *vin* est l'ancêtre, est également ancêtre de *huître* $\rightarrow$ *muscadet*. Par contre, la règle *fruit de mer* $\rightarrow$ *vin blanc* est l'ancêtre proche de *huître* $\rightarrow$ *muscadet*.

Le support attendu d'une règle $a \rightarrow b$ pour une ancêtre $\widehat{a} \rightarrow \widehat{b}$ est :

$$E_{\widehat{a} \cup \widehat{b}}(a \cup b) = \frac{\prod_{x \in a \cup b - \widehat{a} \cup \widehat{b}} support(x)}{\prod_{\widehat{x} \in \widehat{a} \cup \widehat{b} - a \cup b} support(\widehat{x})} \times support(\widehat{a} \cup \widehat{b})$$

Exemple. Considérons les supports suivants : $support(fruit\ de\ mer) = 0,6$, $support(huître) = 0,3$, $support(vin\ blanc) = 0,8$, $support(muscadet) = 0,4$, $support(\{fruit\ de\ mer, vin\ blanc\}) = 0,4$ et $support(\{huître, muscadet\}) = 0,3$, et les deux règles *fruit de mer* $\rightarrow$ *vin blanc* et *huître* $\rightarrow$ *muscadet*. Le support attendu de l'itemset $\{huître, muscadet\}$ par rapport à l'itemset $\{fruit\ de\ mer, vin\ blanc\}$ est $E_{\{fruit\ de\ mer, vin\ blanc\}}(\{huître, muscadet\}) = \frac{support(huître) \times support(muscadet)}{support(fruit\ de\ mer) \times support(vin\ blanc)} \times support(\{fruit\ de\ mer, vin\ blanc\}) = 0,1$.

De manière similaire, la confiance attendue d'une règle $a \rightarrow b$ pour un ancêtre $\widehat{a} \rightarrow \widehat{b}$ est :

$$E_{\widehat{a} \rightarrow \widehat{b}}(a \rightarrow b) = \frac{\prod_{b_i \in b - \widehat{b}} support(b_i)}{\prod_{\widehat{b_i} \in \widehat{b} - b} support(\widehat{b_i})} \times confiance(\widehat{a} \rightarrow \widehat{b})$$

Exemple. A partir de l'exemple précédent, la confiance attendue de la règle *huître* $\rightarrow$ *muscadet* par rapport à la règle *fruit de mer* $\rightarrow$ *vin blanc* est $E_{fruit\ de\ mer \rightarrow muscadet}(huître \rightarrow muscadet) = \frac{support(muscadet)}{support(vin\ blanc)} \times confiance(fruit\ de\ mer \rightarrow vin\ blanc) = 1/3$ ($confiance(fruit\ de\ mer \rightarrow vin\ blanc) = support(\{fruit\ de\ mer, vin\ blanc\})/support(fruit\ de\ mer) = 2/3$).

Règles R-intéressantes [SA95]

Une règle $a \rightarrow b$ sera R-intéressante vis-à-vis d'un de ses ancêtres $\widehat{a} \rightarrow \widehat{b}$ si $support(a \rightarrow b) = R \times E_{\widehat{a} \cup \widehat{b}}(a \cup b)$ ou si $confiance(a \rightarrow b) = R \times E_{\widehat{a} \rightarrow \widehat{b}}(a \rightarrow b)$.

Etant donné un ensemble de règles S et un seuil minimum d'intérê t R, une règle $a \rightarrow b$ sera intéressante dans S si elle n'a pas d'ancêtre ou si elle est R-intéressante vis-à-vis de ses ancêtres proches. Une règle $a \rightarrow b$ sera partiellement intéressante dans S si elle n'a pas d'ancêtre ou si elle est R-intéressante pour au moins un de ses ancêtres proches.

Règles redondantes et inutiles[HF99]

Une règle $a \rightarrow b$ sera redondante vis-à-vis d'un de ses ancêtres $\widehat{a} \rightarrow \widehat{b}$ si $E_{\widehat{a} \rightarrow \widehat{b}}(a \rightarrow b) - \alpha \leq confiance(a \rightarrow b) \leq E_{\widehat{a} \rightarrow \widehat{b}}(a \rightarrow b) + \alpha$.

Une règle $c \rightarrow b$ sera inutile (ou pas nécessaire) s'il existe une règle $a \rightarrow b$, avec $c \subset a$ et $confiance(a \rightarrow b) - \beta \leq confiance(c \rightarrow b) \leq confiance(a \rightarrow b) + \beta$

Avantages et limites

Les deux approches présentées considèrent qu'une règle $a' \rightarrow b'$ n'est pas intéressante (resp. redondante) par rapport à une règle $a \rightarrow b$, avec $a' \leq a$ et $b' \leq b$, si elle n'apporte pas un certain gain de qualité (resp. si elle n'a pas une valeur de confiance très différente). L'avantage principal de ce type d'approche réside dans la comparaison des valeurs de qualité (limitées au support et à la confiance) obtenues par une règle par rapport à celles attendues (sous hypothèse de réduction indépendante et proportionnelle des cardinalités).

La première limite de cette approche concerne une prise en compte partielle de la généralisaton/spécialisation d'itemsets. En effet, la définition d'ancêtre d'un itemset se restreint à la généralisation d'une ou plusieurs de leurs variables et ne prend pas en compte l'inclusion de leurs ensembles de variables.

La deuxième limite concerne l'utilisation de la définition d'ancêtre d'une règle pour réduire la quantité de règles extraites. Par exemple, si l'on considère les règles *huître* $\rightarrow$ *muscadet* et *huître* $\rightarrow$ *vin blanc*, la règle *huître* $\rightarrow$ *muscadet*, qui a un ancêtre *huître* $\rightarrow$ *vin blanc*, sera considérée comme redondante si elle ne répond pas aux critères de qualité utilisés. Cependant, en toute logique, on devrait considérer l'inverse car *huître* $\rightarrow$ *vin blanc* peut être déduite à partir de *huître* $\rightarrow$ *muscadet* étant donné que le muscadet est un vin blanc. De plus, en utilisant ce type d'approche on ne peut pas détecter que *huître* $\rightarrow$ *vin blanc* est redondante par rapport à *fruit de mer* $\rightarrow$ *muscadet* car aucune des deux règles n'est un ancêtre de l'autre.

1.3.2 Réduction des redondances

Dans [PTB+05], les auteurs proposent un autre moyen de réduire la redondance en s'appuyant sur les treillis de galois pour l'extraction des règles. Leur

1- lames libres → comestible
2- lames libres → comestible ∧ voile partiel
3- lames libres → comestible ∧ voile blanc
4- lames libres → comestible ∧ voile partiel ∧ voile blanc
5- lames libres ∧ voile partiel → comestible
6- lames libres ∧ voile partiel → comestible ∧ voile blanc
7- lames libres ∧ voile blanc → comestible
8- lames libres ∧ voile blanc → comestible ∧ voile partiel
9- lames libres ∧ voile partiel ∧ voile blanc → comestible

TAB. 1.3 – Règles redondantes dans une base de données mycologique

définition de la redondance n'est pas la même que [HF99], et le principe suivi n'est pas celui des règles généralisées ([SA95], [HF99]).

Pour illustrer cette notion de redondance, N. Pasquier et al. [PTB+05] prennent un exemple de règles extraites d'une base de données mycologique bien connue : la base de données MUSHROOMS. Sur cet exemple, présenté table 1.3, 9 règles possèdent les mêmes valeurs de confiance et de support et ont toutes la variable « lames libres » dans la prémisse. Parmi cet ensemble de règles, seulement la numéro 4 est intéressante car les autres lui sont redondantes. En effet, la règle numéro 4 a la même confiance et le même support que les autres règles, et de plus, elle possède la prémisse minimale et la conclusion maximale parmi les 9 règles.

A partir de cette notion de redondance, une règle est appelée **règle min-max**, parmi un ensemble de règles S, si elle est non-redondante (c.-à-d. qu'il n'existe pas de règles dans S ayant une prémisse incluse dans la sienne et de conclusion dont la sienne est un sous-ensemble) ou si sa confiance et son support sont égaux à ses règles redondantes.

Définition 1.12 *une règle d'association $a \rightarrow b$ de S est une* ***règle min-max*** *si il n'existe pas de règle $a' \rightarrow b'$ ($\neq a \rightarrow b$) appartenant à S vérifiant toutes les conditions suivantes :*

- $a \subseteq a'$ *et* $b' \subseteq b$
- $confiance(a \rightarrow b) = confiance(a' \rightarrow b')$
- $support(a \rightarrow b) = support(a' \rightarrow b')$

Cette approche a le mérite, par rapport à [SA95], de proposer une réduction des règles extraites sur la base d'une sémantique plus pertinente : « une règle r est redondante par rapport à une règle r' si r a une prémisse plus spécifique ou une conclusion plus générale que r' ». Cependant, elle ne prend pas en compte une éventuelle taxonomie (relation de généralisation/spécialisation) sur les variables.

Une autre limite de cette approche concerne la définition de la règle min-max. Une règle r qui est redondante par rapport à une règle r' sera tout de même conservée si leurs valeurs de support ou de confiance diffèrent. Cependant, lorsque l'on spécialise la conclusion ou que l'on généralise la prémisse, il y a de grande chances que les règles r et r' aient des valeurs de qualité (support et confiance) différentes sans pour autant remettre en cause la redondance et l'inutilité de r par rapport à r'. Par exemple, considérons les règles $fumer \rightarrow$

cancer, $fumer \wedge homme \rightarrow cancer$ et $fumer \wedge femme \rightarrow cancer$ ayant toutes les trois une confiance de 50%. Les règles $fumer \wedge homme \rightarrow cancer$ et $fumer \wedge femme \rightarrow cancer$ auront un support de moitié par rapport à $fumer \rightarrow cancer$ (en considérant qu'il y a autant de femmes que d'hommes). En suivant le critère min-max, les trois règles sont conservées (leurs valeurs de support diffèrent) alors que les deux dernières sont redondantes par rapport à $fumer \rightarrow cancer$ et n'apportent aucune information. Si l'on considère en plus la règle $fumer \rightarrow cancer \wedge tousser$ avec une confiance de 45%, les deux règles $fumer \rightarrow cancer \wedge tousser$ et $fumer \rightarrow cancer$ seront conservées (parce qu'elles ont des confiances différentes) alors que la première règle permet de déduire la seconde et la seconde (même si elle a une confiance supérieure) n'apporte pas vraiment d'information supplémentaire.

1.3.3 Combinaison des deux approches

Nous avons étudié deux principes de réduction du nombre de règles. Chaque approche présente des avantages pour une fouille de règles entre hiérarchies (voir section 2.3) : la première permet de prendre en compte des taxonomies sur les variables ; la deuxième utilise un critère de réduction de redondance plus adapté que la première approche.

A partir des deux approches, nous dirons qu'une règle r est redondante par rapport à une règle r' si r' a une prémisse plus générale et une conclusion plus spécifique que r. Cette définition généralise à la fois l'inclusion entre itemset et l'utilisation d'une taxonomie sur les variables en utilisant la définition 1.7. Ainsi, formellement, $a \rightarrow b$ est redondante par rapport à une règle $a' \rightarrow b'$ ($\neq a \rightarrow b$) si $a \leq a'$ et $b' \leq b$.

Exemple. En reprenant la taxonomie sur les items présentée figure 1.5 et une autre variable « *citron* », considérons les règles *huître* $\rightarrow$ *vin blanc* et *fruit de mer* $\rightarrow$ *muscadet* $\wedge$ *citron*. La première règle est redondante par rapport à la seconde, puisque d'une part, *huître* $\leq$ *fruit de mer*, et d'autre part, $\{muscadet, citron\} \leq vin\ blanc$.

Conclusion

Dans ce chapitre, nous avons tout d'abord présenté le modèle des règles d'association. Nous nous sommes ensuite intéressés plus particulièrement aux mesures d'intérêt utilisées pour juger la qualités des règles d'association. Ces mesures peuvent être classées selon certaines de leur propriétés telles que leur portée (règle, quasi-implication, quasi-conjonction, quasi-équivalence), leur sujet (écart à l'équilibre ou à l'indépendance) et leur nature (statistique ou descriptive).

Dans un deuxième temps, nous avons étudié deux approches principales de réduction de la redondance : les règles généralisées et l'approche de [PTB+05] (basée sur l'axiome d'augmentation d'Amstrong). Ces deux approches peuvent être combinées dans le but de généraliser la notion de redondance à l'utilisation de taxonomies d'items et à l'inclusion entre itemsets.

Modèles de hiérarchie et d'alignement

2

Sommaire

Introduction

Dans les systèmes d'information et sur le Web, les données et les connaissances, souvent sous forme textuelle, ont tendance à être structurées en hiérarchies. Ces représentations hiérarchiques regroupent des structures peu formalisées, comme les systèmes de fichiers, en passant par des représentations semi-formelles, comme les thésaurus, jusqu'aux théories logiques, telles que les ontologies. Le premier objectif de ce chapitre est de présenter un modèle de hiérarchie assez général pour prendre en compte une grande variété de représentations.

Dans un second temps, nous introduisons notre modèle d'alignement entre hiérarchies. Ce modèle étend en partie les formalismes précédemment introduits dans la littérature [SEN+06], [RB01]. En effet, nous insistons, plus particulièrement, sur les aspects implicatifs d'un alignement. A partir de règles de déductions, nous étudions les notions de fermeture et couverture minimale d'un alignement. Ces notions nous permettent ensuite d'expliquer la redondance d'un alignement (engendrée par la prise en compte de la relation d'implication). Nous étudions également la consistance, et les notions de symétrie et de cardinalité d'un alignement.

Finalement, nous introduisons notre modèle de règles d'association entre hiérarchies.

2.1 Modèle de hiérarchie conceptuelle

Nous distinguons et formalisons deux modèles de hiérarchies. Le premier modèle est celui de la hiérarchie hors-contexte. Il représente le modèle de base utilisé pour la description de schémas objets, XML, ou encore d'ontologies. Le deuxième modèle, appelé hiérarchie contextualisée, est une extension du premier. Une telle hiérarchie possède les mêmes éléments qu'une hiérarchie hors-contexte (ce qui constitue sa définition dite intensionnelle) et possède en plus une extension constituée d'un ensemble d'objets qui seront indexés aux entités de la hiérarchie.

Avant de définir les modèles de hiérarchie hors-contexte et contextualisée, nous présentons quelques rappels sur les ensembles ordonnés, qui font partie intégrante d'une hiérarchie.

2.1.1 Rappels sur les ensembles ordonnés

Définition 2.1 *Une relation binaire R entre deux ensembles M et N est un ensemble de couples $(m, n) \in M \times N$. Un couple $(m, n) \in R$ est noté mRn. Si $M = N$ alors R est une relation binaire sur l'ensemble M. R^{-1} représente la relation inverse de R : $mRn \Leftrightarrow nR^{-1}m$.*

Définition 2.2 *Une relation binaire R sur un ensemble M est appelée relation d'ordre, si pour tout élément $x, y, z \in M$, elle est :*

1. *Réflexive : xRx*
2. *Antisymétrique : $xRy \wedge x \neq y \Rightarrow \neg(yRx)$*
3. *Transitive : $xRy \wedge yRz \Rightarrow xRz$*

Une relation d'ordre R est souvent désignée par le symbole $\leq$ ($\geq$ pour R^{-1}) et est lue « x est plus petit ou égal à y ». La notation $x < y$ désigne $x \leq y \wedge x \neq y$. Un ensemble ordonné est un couple $(M, \leq)$ où M est un ensemble et $\leq$ est une relation d'ordre sur M.

Définition 2.3 *a est un prédécesseur direct (ou fils) de b, si $a < b$ et s'il n'existe pas d'élément c satisfaisant $a < c < b$. Dans ce cas, nous dirons également que*

b est un successeur direct (ou père) de a et nous utiliserons la notation $a \prec b$ ou encore $b \succ a$.

Définition 2.4 *Soit $(M, \leq)$ un ensemble ordonné, A une partie de M et $x \in M$.*
- *x est un majorant de A si $\forall a \in A \ a \leq x$.*
- *x est un minorant de A si $\forall a \in A \ x \leq a$.*
- *x est le plus grand élément de A si $x \in A$ et $\forall a \in A \ a \leq x$.*
- *x est le plus petit élément de A si $x \in A$ et $\forall a \in A \ x \leq a$.*

2.1.2 Hiérarchie hors-contexte

Une hiérarchie hors-contexte, ou non peuplée, est une structure permettant de définir un ensemble d'entités et de les organiser en hiérarchie par une relation d'ordre partiel. Une telle hiérarchie possède également des fonctions d'annotation qui permettent d'associer des descriptions textuelles à chaque entité.

Définition 2.5 *Une hiérarchie hors-contexte est définie par le triplet :*

$$\mathcal{H} = (C, \leq, \mathcal{A})$$

où :
- *C représente l'ensemble des entités.*
- *$\leq$ représente la relation d'ordre partiel entre les entités. Ainsi le couple $(C, \leq)$ est un ensemble ordonné.*
- *$\mathcal{A}$ regroupe les fonctions d'annotations $\mathcal{A}_x$ associant une description textuelle aux entités.*

L'ensemble ordonné $(C, \leq)$ d'une hiérarchie possède un plus grand élément qui est appelé racine.

Les entités dénotent, selon le contexte d'utilisation, différents types d'objets. Dans le cas d'annuaires de sites Web, ou de catalogues de boutiques en ligne, les entités seront des catégories ou rubriques. Pour les ontologies, ces entités pourront représenter soit des classes (concepts), soit des propriétés.

Plusieurs sémantiques peuvent être associées à la relation d'ordre partiel. Cette relation d'ordre partiel peut être une relation de spécialisation (relation *est un*), de composition (relation *partie de*). La sémantique associée dépend du type de hiérarchie considéré. Dans le cas de schémas objets, d'ontologies ou de taxonomies, la relation sera une spécialisation. Si l'on s'intéresse à la structure hiérarchique (physique) d'un document XML, le type de relation sera une composition. Cependant au sein d'une même hiérarchie, la sémantique associée à la relation d'ordre devra être unique. En effet, si l'on combine plusieurs sémantiques pour une même relation d'ordre, la propriété de transitivité ne sera plus assurée. Par exemple, à partir d'une hiérarchie définissant *voiture* $\prec$ *véhicule* , *vélo* $\prec$ *véhicule* (relations est un) et *moteur* $\prec$ *voiture* (relation partie de), la relation (déduite par transitivité) *moteur* $\leq$ *véhicule* n'a pas de sens. Un moteur n'est pas un véhicule et un moteur n'est pas forcément une partie d'un véhicule (un vélo n'a pas de moteur).

Les fonctions d'annotations regroupées par $\mathcal{A}$ servent à nommer et décrire les entités en langage naturel. On peut définir une multitude de fonctions d'annotation qui permettent, par exemple, d'associer à une entité son identifiant,

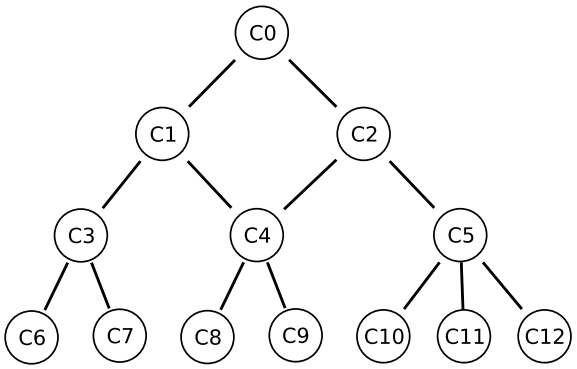


FIG. 2.1 – Représentation graphique d'une hiérarchie

des noms (ou labels), des commentaires etc. Dans le cadre de cette thèse, nous distinguons 3 fonctions d'annotations principales :

- $\mathcal{A}_{id}$ associant un identifiant à chaque entité $c_i \in C$. Cette fonction d'annotation est la seule qui est toujours présente.
- $\mathcal{A}_{label}$ associant un ensemble de labels à une entité. Ces labels permettent de nommer une entité de manière plus explicite que l'identifiant. Une entité peut avoir plusieurs labels.
- $\mathcal{A}_{com}$ associant un ensemble de commentaires à une entité. Un commentaire peut être, par exemple, une définition de l'entité en langage naturel.

Pour une entité $c_i \in C$, nous utiliserons la notation $\mathcal{A}_x(c_i)$ (avec $x = \{id, label, com\}$) pour dénoter les annotations de type identifiant, label, ou commentaire. La notation $\mathcal{A}(c_i)$ désigne l'ensemble des annotations associées à c_i, indépendamment de leur type.

Représentation graphique

Une hiérarchie est représentée graphiquement par un diagramme comme celui présenté figure 2.1 (appelé diagramme de Hasse), où chaque entité est symbolisée par une ellipse (ou également un rectangle) contenant son identifiant (donné par la fonction d'annotation $\mathcal{A}_{id}$). La relation d'ordre est symbolisée d'une part, par la position des entités : si $c_i \leq c_j$ alors la représentation de l'entité c_j sera placée plus haut que celle de c_i, et d'autre part, par sa réduction transitive et réflexive : si $c_i \prec c_j$ alors un segment relie la représentation de c_i à celle de c_j.

Exemple. Considérons une hiérarchie sur les véhicules représentée figure 2.2. L'ensemble des entités est $C = \{$*véhicule, véhicule motorisé, véhicule non motorisé, véhicule léger, poids lourd, voiture, moto, camion, car, véhicule tracté, véhicule à vent, remorque, caravane, voilier, planche à voile*$\}$. La relation d'ordre partiel est une relation *est un* : Une « *voiture* » est un « *véhicule* ». A partir de cette relation, on peut déduire, par exemple, *voiture* $\leq$ *véhicule* ou encore *voiture* $\prec$ *véhicule léger*. Dans cette hiérarchie, chaque entité est désignée grâce à la fonction d'annotation $\mathcal{A}_{id}$. Un autre fonction d'annotation $\mathcal{A}_{label}$ appliquée

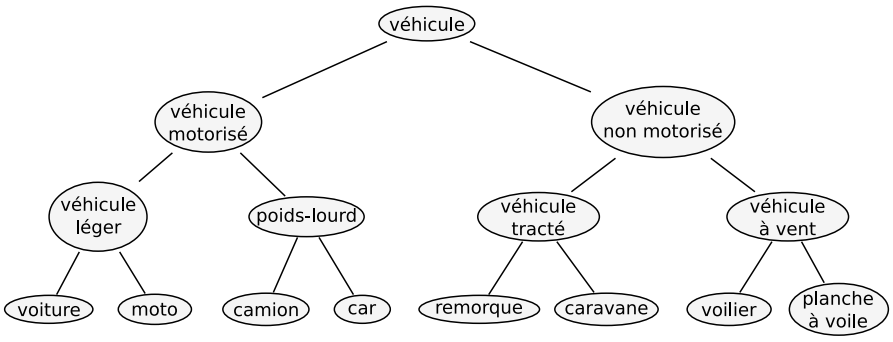


FIG. 2.2 – Exemple d'une hiérarchie sur les véhicules

sur l'entité *"voiture"* pourrait retourner l'ensemble de ses dénominations : $\mathcal{A}_{label}(voiture) = \{voiture, automobile, caisse\}$.

Constituants optionnels d'une hiérarchie

Une hiérarchie peut être également pourvue d'un ensemble de relations, notées $\mathcal{P}$. Il existe deux types de relations possibles :

- $\mathcal{P}_c \subseteq C \times C$: l'ensemble des relations transversales inter-entités.
- $\mathcal{P}_a \subseteq C \times A$: l'ensemble des propriétés (ou attributs) d'une entité. A désigne un attribut défini sur un type de donnée simple : date, chaîne de caractères, entier, réel, etc.

Exemple. Reprenons la hiérarchie sur les véhicules (figure 2.2). A tout type de *véhicule motorisé*, on peut associer, par exemple, les attributs « puissance fiscale », « cylindrée », « type d'énergie ». On peut également ajouter une relation transversale « tracter » entre *voiture* et *véhicule tracté*, dénotant le fait qu'une voiture peut tracter un *véhicule tracté*.

Remarque. Etant donné que la sémantique associée à la relation d'ordre est la spécialisation (*est un*), on est capable également de déduire qu'une *voiture* peut tracter une *remorque* ou une *caravane*. Cependant, il est à noter que cette déduction ne pourrait pas être faite dans le cas d'une relation d'ordre ayant une sémantique de composition. Par exemple, si l'on considère qu'un *véhicule tracté* est composé de *roues* (*roue* $\leq$ *véhicule tracté*), on ne peut pas déduire qu'une *voiture* tracte des *roues*. Cette remarque est également applicable aux propriétés : les propriétés associées à une entité ne peuvent être héritées à ses descendantes que si la sémantique associée à relation d'ordre partiel est la spécialisation. Par exemple, si un véhicule tracté a un attribut *immatriculation*, son composant *roue* n'hérite pas de l'attribut *immatriculation*.

2.1.3 Hiérarchie contextualisée

Une hiérarchie contextualisée (appelée également hiérarchie peuplée ou intanciée) est une hiérarchie hors-contexte à laquelle on ajoute un ensemble d'objets (ou instances) qui seront associés aux entités.

Définition 2.6 *Une hiérarchie contextualisée est définie par :*

$$\mathcal{H} = (C, \leq, \mathcal{A}, O, \sigma)$$

où :

- *O représente l'ensemble des objets peuplant la hiérarchie.*
- *$\sigma : C \longrightarrow 2^O$ est la relation d'association (également appelée relation d'indexation) des entités aux objets. Pour tout $c_i \in C$, $\sigma(c_i)$ représente les objets associés à l'entité c_i.*

Propriétés :

- *$c_i \leq c_j$ si et seulement si $\sigma(c_i) \subseteq \sigma(c_j)$. σ est un isomorphisme de $(C, \leq)$ dans $(2^O, \subseteq)$.*
- *$\sigma(c_0) = O$: l'ensemble des objets est associé à l'entité racine.*

Les objets associés à une hiérarchie sont des instances dans le cas d'ontologies ou de schémas objets, des documents textuels dans le cas d'index hiérarchiques du type thésaurus, de catalogues de boutiques en lignes, ou de répertoires de sites Web.

La partie hors-contexte d'une hiérarchie contextualisée sera par la suite appelée intension de la hiérarchie. Lorsque l'on s'intéressera également à l'ensemble des objets et à la relation d'association, nous parlerons alors d'extension de la hiérarchie.

Exemple. La figure 2.3 reprend le thème des véhicules. Cette fois, les véhicules sont classés en fonction du médium de transport (eau ou terre). Cette hiérarchie est contextualisée par un ensemble d'objets O qui sont, par exemple, des descriptifs des produits vendus par un marchand de véhicules. Comme *voiture* $\leq$ *véhicule terrestre*, l'ensemble des objets associés à *voiture*, $\sigma(voiture) = \{o_1, o_2\}$ est inclus dans l'ensemble des objets associés à l'entité *véhicule terrestre*. L'ensemble d'objets associés à l'entité *aéroglisseur*, $\sigma(aéroglisseur) = \{o_3\}$, est à la fois inclus dans ceux de *véhicule terrestre* et *véhicule nautique*. Comme les entités *véhicule terrestre* et *véhicule nautique* partagent un prédécesseur en commun, l'intersection $\sigma(véhicule\ terrestre) \cap \sigma(véhicule\ nautique)$ n'est pas vide. L'entité racine, *véhicule*, est quant à elle, associée à l'ensemble des objets O par σ.

2.2 Modèle d'alignement

Dans le cadre des représentations hiérarchiques, nous définissons un alignement entre deux hiérarchies comme un ensemble de relations entre entités issues des deux structures. Nous ne nous limitons pas seulement à identifier de simples appariements entre entités, mais nous prenons en compte la nature de l'appariement en distinguant, notamment, les relations d'équivalence et d'implication entre entités.

Nous présentons, dans cette section, notre définition formelle d'un alignement entre deux hiérarchies. Nous mettons également l'accent sur la notion de redondance au sein d'un alignement qui peut apparaître lorsque l'on prend en compte la relation d'implication. Nous expliquerons également la notion de consistance d'un alignement.

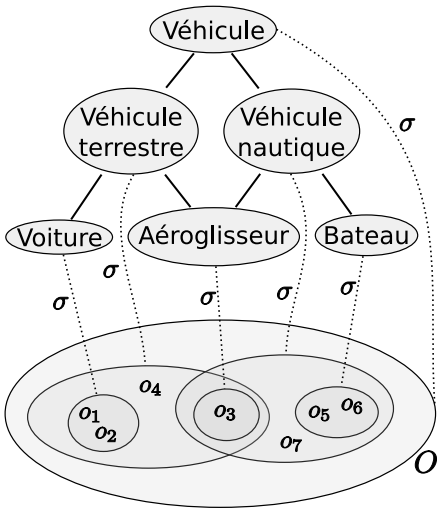


FIG. 2.3 – Exemple de hiérarchie peuplée

2.2.1 Définition d'un alignement

Définition 2.7 *Un alignement A entre deux hiérarchies $\mathcal{H}_1 = (C_1, \leq, \mathcal{A}_1)$ et $\mathcal{H}_2 = (C_2, \leq, \mathcal{A}_2)$ est un couple (V, q) où :*

- *V représente un ensemble d'éléments de correspondances désignés par un triplet $a = (x, y, \mathcal{R})$ où :*
  - *$x \in C_1$ et $y \in C_2$.*
  - *$\mathcal{R}$ est la relation qu'entretiennent les entités x et y. Cette relation peut être une implication ($\Rightarrow$ ou $\Leftarrow$) ou une équivalence ($\Leftrightarrow$).*
- *$q : V \longrightarrow \mathbb{R}$ est une application associant à chaque élement $a \in V$ une valeur quantifiant la qualité de la relation $\mathcal{R}$ entretenue entre x et y.*

Un élément de correspondance $a = (x, y, \mathcal{R})$ pourra être également noté $x\mathcal{R}y$ (exemples : $x \Rightarrow y$ ou $y \Leftrightarrow y$).

Remarque. Afin que l'alignement soit interprétable, les relations d'ordre respectives des deux hiérarchies $\mathcal{H}_1$ et $\mathcal{H}_2$ doivent avoir une sémantique identique (soit deux relations *est-un* ou *partie-de*).

L'ensemble V peut être ramené à l'union des deux relations d'implication. En effet, la relation d'équivalence entre deux entités est déduite à partir de la composition de deux implications $\Leftarrow$ et $\Leftarrow$. Ainsi une équivalence $x \Leftrightarrow y$ correspond à la fusion des deux élements de correspondance $x \Rightarrow y$ et $x \Leftarrow y$.

Contrairement à la définition d'alignement proposée par [SE05], notre définition de l'alignement sépare l'ensemble des éléments de correspondance V et la fonction de qualité q. Cette séparation, nous permet de considérer l'ensemble V comme l'union de deux relations $\Rightarrow$ et $\Leftarrow$ ($\Leftrightarrow$ pouvant être déduite à partir des deux premières). Nous pouvons ainsi utiliser les propriétés de ces relations d'implication. Cela nous permet par la suite de définir formellement des notions telles que l'égalité (et inclusion) entre deux alignements (qui ne prend

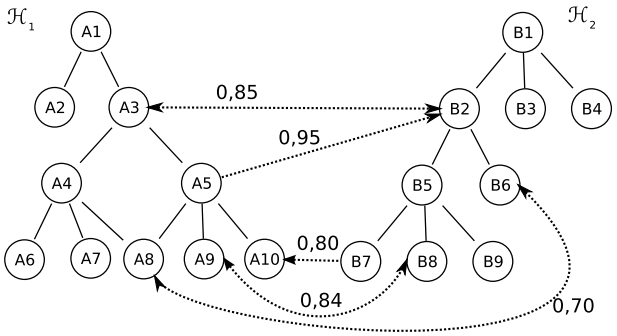


FIG. 2.4 – Représentation graphique d'un alignement

pas en compte les valeurs de qualité associés aux éléments de correspondance), la redondance dans un alignement ou la déduction d'équivalences à partir d'implications.

Représentation graphique d'un alignement

La représentation graphique d'un alignement $A(V, q)$ de deux hiérarchies $\mathcal{H}_1$ et $\mathcal{H}_2$ est illustrée sur la figure 2.4. Elle est composée des représentations des deux hiérarchies $\mathcal{H}_1$ et $\mathcal{H}_2$ et d'un ensemble de liens (en pointillés) entre des entités issues respectivement des deux hiérarchies. Un lien pourvu d'une flèche à une seule de ses extrémités représente une implication : par exemple $A5 \Rightarrow B2$. Un lien ayant une flèche à chacune de ses extrémités représente une équivalence : par exemple $A3 \Leftrightarrow B2$. Les nombres associés aux liens représentent la valeur de qualité donnée par la fonction q à l'élément de correspondance. Par exemple la qualité de l'élément $A5 \Rightarrow B2$ est $q(A5 \Rightarrow B2) = 0,95$

L'alignement $A(V, q)$ représenté contient l'ensemble des éléments de correspondance $V = \{A3 \Leftrightarrow B2, A5 \Rightarrow B2, A8 \Leftrightarrow B6, A9 \Leftrightarrow B8, A10 \Leftarrow B7\}$, associé à la fonction de qualité $q = \{(A3 \Leftrightarrow B2; 0,85), (A5 \Rightarrow B2; 0,95), (A8 \Leftrightarrow B6; 0,70), (A9 \Leftrightarrow B8; 0,84), (A10 \Leftarrow B7; 0,80)\}$

Sémantique des relations considérées dans un alignement

Un alignement peut contenir deux types de relations : l'implication et l'équivalence. Une équivalence entre deux entités x et y signifie que ces entités sont considérées comme identiques (avec éventuellement un niveau de confiance donné par $q(x \Leftrightarrow y)$). La signification d'une implication entre deux entités x et y dépend de la sémantique des relations d'ordre partiel associées aux hiérarchies. Dans le cas d'une relation de spécialisation, *est un*, l'implication $x \Rightarrow y$ signifiera que l'entité x est une sorte d'entité y. Dans le cas d'une relation de composition, *partie de*, l'implication signifiera que l'entité x est une partie de l'entité y.

La sémantique associée à l'implication est la même que celle de la relation d'ordre partiel. L'équivalence a le même sens que l'égalité. Nous n'utilisons

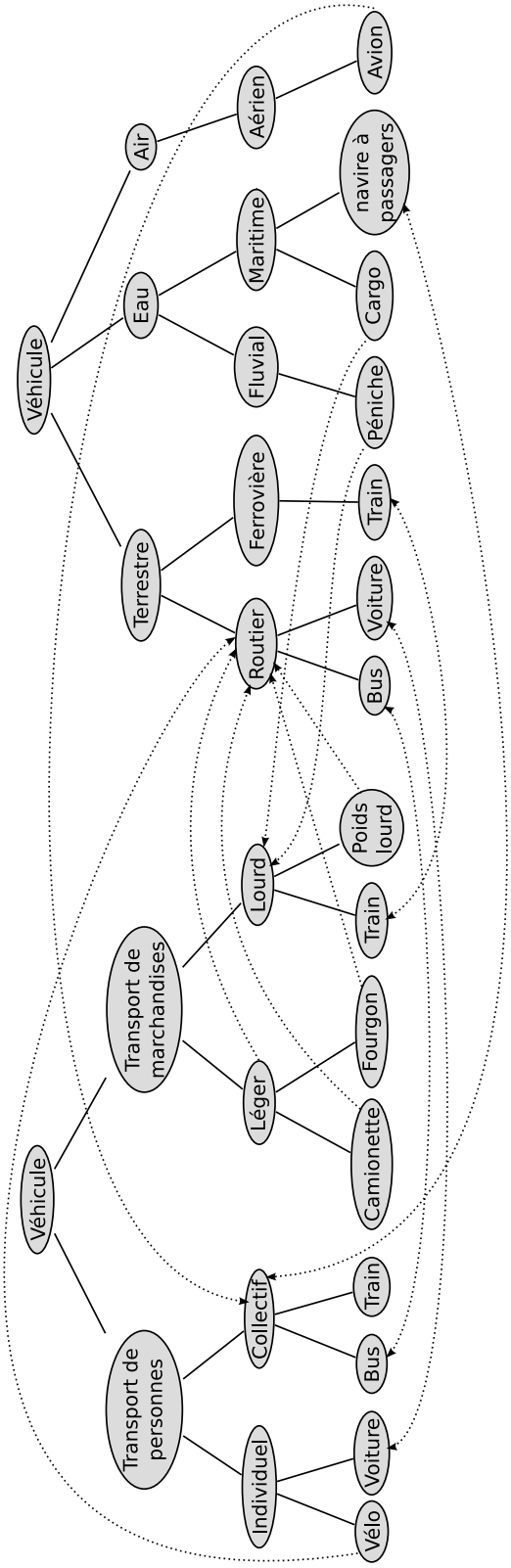


FIG. 2.5 – Exemple d'alignement entre deux hiérarchies de véhicules

cependant pas les mêmes notations : $\leq$ pour les relations d'ordre partiel et $\Rightarrow$ pour l'implication dans l'alignement. Cela nous permet de distinguer clairement les relations relevant des hiérarchies de celles contenues dans l'alignement.

Par exemple, sur la figure 2.5, la relation *Léger* $\Rightarrow$ *Routier* est interprétée : « Un véhicule de transport de marchandise léger est un véhicule routier ».

D'un point de vue extensionnel et en considérant que les deux hiérarchies partagent le même ensemble d'objets $O_1 = O_2$, une relation d'implication $x \Rightarrow y$ représentera une inclusion de l'ensemble des objets associés à l'entité x dans l'ensemble des objets associés à y : $\sigma_1(x) \subseteq \sigma_2(y)$.

2.2.2 Déduction à partir d'un alignement

Il est possible d'associer des règles d'inférence permettant à partir d'un alignement et des hiérarchies associées, de déduire, s'il en existe, de nouveaux éléments de correspondance. Pour cela, nous nous basons sur le postulat que les relations d'ordre partiel des hiérarchies $\leq$ et la relation d'implication $\Rightarrow$ ont une même sémantique, compatible avec la logique des propositions.

Déduction par transitivité

Soit $A = (V, q)$ un alignement entre deux hiérarchies $\mathcal{H}_1$ et $\mathcal{H}_2$. Soit les entités $x, x' \in C_1$ et $y, y' \in C_2$. En confondant les relations d'ordre partiel et l'implication, et en utilisant la transitivité, nous obtenons les deux règles suivantes :

- Si $x' \leq x$ et $x \Rightarrow y$ alors $x' \Rightarrow y$
- Si $y \leq y'$ et $x \Rightarrow y$ alors $x \Rightarrow y'$

Ces règles ont été énoncées pour des implications d'entités de c_1 vers des entités de c_2. Elles sont naturellement valables pour des implications de sens inverse.

Du côté extensionnel, en supposant que les hiérarchies partagent les mêmes extensions, $O_1 = O_2$, ces règles traduisent la transitivité de la relation d'inclusion :

- $x \leq x'$ et $x' \Rightarrow y \models \sigma(x) \subseteq \sigma(x')$ et $\sigma(x') \subseteq \sigma(y) \models \sigma(x) \subseteq \sigma(y)$
- $y \leq y'$ et $x \Rightarrow y \models \sigma(y) \subseteq \sigma(y')$ et $\sigma(x) \subseteq \sigma(y) \models \sigma(x) \subseteq \sigma(y')$

Exemple. En prenant l'alignement illustré figure 2.5, nous avons l'implication *Vélo* $\Rightarrow$ *Routier* et *Routier* $\leq$ *Terrestre*. Nous pouvons alors déduire (grâce à la deuxième règle) *Vélo* $\Rightarrow$ *Terrestre*, signifiant qu'un *vélo* est un *véhicule terrestre*. La première règle permet, par exemple, de déduire à partir de *Léger* $\Rightarrow$ *Routier* et de *Fourgon* $\leq$ *Léger*, l'implication *Fourgon* $\Rightarrow$ *Routier*. Cette implication signifie qu'un *fourgon* est un *véhicule routier*.

Fermeture et réduction d'un alignement

A partir des règles de déduction, nous pouvons ensuite définir la fermeture transitive et la réduction transitive d'un alignement.

Définition 2.8 *La* ***fermeture transitive*** *(ou clôture) d'un alignement est l'alignement contenant l'ensemble des éléments de correspondance, noté* V^+*,*

qui peuvent être déduits à partir de V en utilisant les hiérarchies et les règles de déductions.

$$V^+ = \{x' \Rightarrow y' | x' \leq x \text{ et } y \leq y' \text{ et } x \Rightarrow y \in V\} \cup \{x' \Leftarrow y' | x \leq x' \text{ et } y' \leq y \text{ et } x \Leftarrow y \in V\} \quad (2.1)$$

Exemple. A partir de l'alignement $A = (V, q)$ présenté sur la figure 2.4, les fermetures de chacune des implications contenues dans V sont données dans la liste suivante. Nous utilisons, pour cela, la notation $\{x_1, ..., x_n\} \Rightarrow \{y_1, ..., y_m\}$ représentant l'ensemble des implications possibles entre les éléments des couples issus du produit cartésien $\{x_1, ..., x_n\} \times \{y_1, ..., y_m\}$. Par exemple, la notation $\{x_1, x_2\} \Rightarrow \{y_1, y_2\}$ regroupe les implications $x_1 \Rightarrow y_1$, $x_1 \Rightarrow y_2$, $x_2 \Rightarrow y_1$, $x_2 \Rightarrow y_2$.

- $A3 \Rightarrow B2 : \{A3, A4, A5, A6, A7, A8, A9, A10\} \Rightarrow \{B1, B2\}$
- $A8 \Rightarrow B6 : \{A8\} \Rightarrow \{B1, B2, B6\}$
- $A9 \Rightarrow B8 : \{A9\} \Rightarrow \{B1, B2, B5, B8\}$
- $A3 \Leftarrow B2 : \{A3, A1\} \Leftarrow \{B2, B5, B6, B7, B8, B9\}$
- $A8 \Leftarrow B6 : \{A1, A3, A4, A5, A8\} \Leftarrow \{B6\}$
- $A9 \Leftarrow B8 : \{A1, A3, A5, A9\} \Leftarrow \{B8\}$
- $A10 \Leftarrow B7 : \{A1, A3, A5, A10\} \Leftarrow \{B7\}$

La fermeture transitive de V sera alors :

$$V^+ = \big\{\{A3, A4, A5, A6, A7, A8, A9, A10\} \Rightarrow \{B1, B2\}, A8 \Rightarrow B6, A9 \Rightarrow \{B5, B8\}\big\} \cup \big\{\{A3, A1\} \Leftarrow \{B2, B5, B6, B7, B8, B9\}, A5 \Leftarrow \{B6, B8, B7\}, \{A4, A8\} \Leftarrow B6, A9 \Leftarrow B8, A10 \Leftarrow B7\big\}$$

Définition 2.9 *Une **réduction transitive** (ou couverture minimale) d'un alignement est constituée d'un ensemble minimal d'éléments de correspondance, noté V^0, permettant de déduire V.*

$$V^0 = \big\{x \Rightarrow y | \nexists (x', y')\ x' \Rightarrow y' \in V^+ \text{ et } x < x' \text{ et } y' < y\big\} \cup \big\{x \Leftarrow y | \nexists (x', y')\ x' \Leftarrow y' \in V^+ \text{ et } x' < x \text{ et } y < y'\big\} \quad (2.2)$$

Exemple. Une réduction transitive de l'alignement $A = (V, q)$ présenté sur la figure 2.4 est :

$$V^0 = \big\{A3 \Rightarrow B2, A8 \Rightarrow B6, A9 \Rightarrow B8\big\} \cup \big\{A3 \Leftarrow B2, A10 \Leftarrow B7, A9 \Leftarrow B8, A8 \Leftarrow B6\big\}$$

La seule règle de l'alignement initial qui n'apparaît pas dans V^0 est $A5 \Rightarrow B2$. Cette règle peut être déduite à partir de $A3 \Rightarrow B2$ et $A5 \leq A3$.

Egalité et inclusion d'alignements

Nous définissons l'égalité entre deux alignements $A_1 = (V_1, q_1)$ et $A_2 = (V_2, q_2)$, à partir de leurs ensembles de correspondances respectifs. L'idée sous-

jacente est la suivante : deux alignements sont égaux s'ils permettent de déduire les mêmes ensembles de correspondances. Notre définition de l'égalité entre alignements ne prend pas en compte les fonctions de qualité associées.

Définition 2.10 *Les deux alignements A_1 et A_2 seront* ***égaux*** *si les fermetures de leurs ensembles de correspondances respectifs sont égales.*

$$A_1 = A_2 \Longleftrightarrow V_1^+ = V_2^+$$

De la même manière, nous définissons l'inclusion d'un alignement A_1 dans l'alignement A_2 comme l'inclusion de leurs fermetures respectives. Une inclusion d'un alignement A_1 dans un alignement A_2 signifie que l'ensemble des éléments de correspondance déductibles à partir de $A1$ peut être déduit à partir de A_2. En d'autres termes, l'alignement A_2 est plus riche que l'alignement A_1.

Définition 2.11 *Un alignement A_1 est inclus dans A_2 si la fermeture de l'ensemble des correspondances de A_1 est incluse dans celle de A_2*

$$A_1 \subseteq A_2 \Longleftrightarrow V_1^+ \subseteq V_2^+$$

2.2.3 Redondance dans un alignement

Lorsqu'un alignement distingue l'implication (et ne considère pas seulement l'équivalence), des éléments de correspondances peuvent être redondants par rapport à d'autres. Un élément est dit redondant s'il peut être déduit à partir d'un autre élément de correspondance en utilisant les règles de déductions.

Définition 2.12 *Une implication $x \Rightarrow y$ sera* ***redondante*** *par rapport à une implication $x' \Rightarrow y'$ $(\neq (x \Rightarrow y))$ si $x \leq x'$ et $y' \leq y$. Nous dirons également que l'implication $x' \Rightarrow y'$ est* ***génératrice*** *de l'implication $x \Rightarrow y$.*

Exemple. Pour illustrer cette notion de redondance, considérons l'alignement, défini sur la figure 2.6, entre deux hiérarchies $\mathcal{H}_1$ et $\mathcal{H}_2$. Sur cet alignement, les implications $A5 \Rightarrow B2$ et $A10 \Rightarrow B5$ sont redondantes par rapport à $A3 \Rightarrow B5$. En effet, d'une part, comme $A5 \leq A3$ et $B5 \leq B2$, nous pouvons déduire à partir de $A3 \Rightarrow B5$, l'implication $A5 \Rightarrow B2$, et d'autre part, comme $A10 \leq A3$ et $B5 \leq B2$, nous pouvons également déduire $A10 \Rightarrow B5$.

La notion de redondance dans un alignement est intrinsèquement liée à celle de réduction transitive d'un alignement. En effet, un alignement $A = (V, q)$ ne contiendra pas d'implications redondantes si la réduction transitive de son ensemble d'éléments de correspondance est lui-même, c'est-à-dire si $V = V^0$. Dans les autres cas, c.-à-d. lorsque $V \neq V^0$, l'alignement contiendra alors des redondances.

On peut remarquer qu'un élément de correspondance a est redondant dans un ensemble V si la fermeture de V est tout de même égale à la fermeture de V privée de a. Formellement, a est redondant si $(V - \{a\})^+ = V^+$.

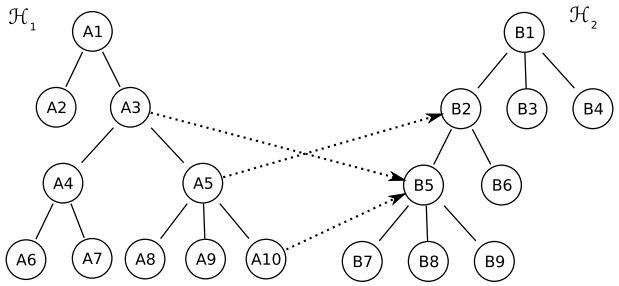


FIG. 2.6 – Relations redondantes dans un alignement

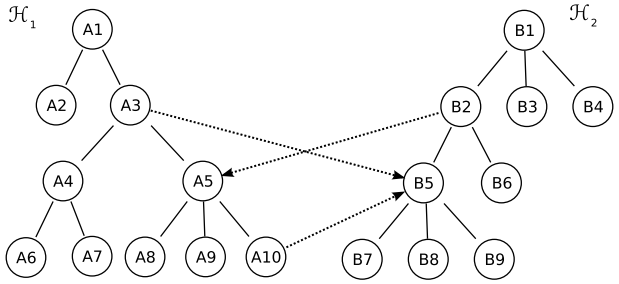


FIG. 2.7 – Inconsistance dans un alignement

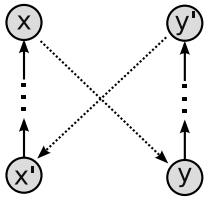
2.2.4 Consistance d'un alignement

La consistance dans un alignement est définie comme l'absence d'éléments de correspondance en contradiction vis-à-vis des relations d'ordre définies sur les hiérarchies. Deux éléments de correspondance sont en contradiction s'ils remettent en cause les connaissances issues des hiérarchies.

Définition 2.13 *Deux éléments de correspondance $x \Rightarrow y$ et $x' \Leftarrow y'$, lorsque $x \neq x'$ ou $y \neq y'$, sont en contradiction si $y \leq y'$ et $x' \leq x$.*

Si l'on pose $x' \leq x \equiv x' \Rightarrow x$ ($y \leq y' \equiv y \Rightarrow y'$) et $x' \Leftarrow y' \equiv y' \Rightarrow x'$, on obtient par transitivité $x \Rightarrow y'$ à partir de $x \Rightarrow y$, $y \Rightarrow y'$, $y' \Rightarrow x'$. Comme $x' \Rightarrow x$, on a alors $x = x'$. Cependant cette équivalence est contradictoire par rapport à la connaissance $x \neq x'$ issue de la hiérarchie.

En considérant un arc orienté de a vers a' lorsque $a \prec a'$, une insconsistance se traduit par un cycle dans le graphe représentant l'alignement :



Exemple. Sur la figure 2.7, les implications $A3 \Rightarrow B5$ et $A5 \Leftarrow B2$ sont en contradiction puisque $A5 \leq A3$ et $B5 \leq B2$. Nous pouvons également remarquer que l'implication $A10 \Rightarrow B5$ qui est redondante par rapport à $A3 \Rightarrow B5$ n'entraîne pas d'inconsistance avec $A5 \Leftarrow B2$.

2.2.5 Symétrie et cardinalité d'un alignement

L'ensemble des éléments de correspondance V d'un alignement A peut être considéré comme un ensemble de relations $(\Rightarrow, \Leftrightarrow, \Leftarrow)$ de l'ensemble C_1 (entités de la hiérarchie source) vers l'ensemble C_2 (entités de la hiérarchie cible). Nous pouvons décomposer cet ensemble V en deux sous-ensembles contenant d'une part, les élements des relations $\Rightarrow$ et $\Leftrightarrow$, et d'autre part les élements des relations $\Leftarrow$ et $\Leftrightarrow$. Ces deux sous-ensembles, appelés composantes (ou restrictions) asymétriques, seront notés respectivement $V_{\Rightarrow}$ et $V_{\Leftarrow}$. L'union de ces deux sous-ensembles sera égale à l'ensemble initial : $V_{\Rightarrow} \cup V_{\Leftarrow} = V$. L'intersection de ces deux ensembles contiendra uniquement les éléments de la relation d'équivalence $\Leftrightarrow$: $V_{\Rightarrow} \cap V_{\Leftarrow} = V_{\Leftrightarrow}$. L'ensemble $V_{\Leftrightarrow}$ est appelé composante fortement symétrique de V. Les ensembles $V_{\Rightarrow} - V_{\Leftrightarrow}$ et $V_{\Leftarrow} - V_{\Leftrightarrow}$ contiennent seulement des implications et sont appelés composantes fortement asymétriques.

Un alignement A sera dit *asymétrique* si son ensemble d'éléments de correspondance V est égal à l'une de ses composantes asymétriques $V_{\Rightarrow}$ ou $V_{\Leftarrow}$. Dans ce cas, l'autre composante asymétrique sera évidemment vide.

Définition 2.14 *Un alignement A est dit **asymétrique** et orienté de $\mathcal{H}_1$ vers $\mathcal{H}_2$ (resp. de $\mathcal{H}_2$ vers $\mathcal{H}_1$) si $V_{\Rightarrow} = V$ (resp. si $V_{\Leftarrow} = V$). Un tel alignement sera noté $A_{\mathcal{H}_1 \Rightarrow \mathcal{H}_2}$ (resp. $A_{\mathcal{H}_2 \Rightarrow \mathcal{H}_1}$).*

Un alignement asymétrique est un *alignement implicatif* s'il est seulement composé d'éléments de correspondance de type implication. Un tel alignement a une composante fortement symétrique vide. Ainsi, l'ensemble de ses éléments de correspondance V sera égal à sa composante fortement asymétrique.

Définition 2.15 *Un alignement A est dit **implicatif** et orienté de $\mathcal{H}_1$ vers $\mathcal{H}_2$ (resp. de $\mathcal{H}_2$ vers $\mathcal{H}_1$) si $V_{\Rightarrow} = V$ (resp. si $V_{\Leftarrow} = V$) et si $V_{\Leftrightarrow} = \emptyset$.*

Exemple. La figure 2.8 montre un alignement symétrique. En effet, dans cet alignement les trois relations sont représentées. Il contient, par exemple, $A3 \Leftrightarrow B2$, $A5 \Rightarrow B2$ et $A10 \Leftarrow B7$. La figure 2.9 présente ses deux composantes asymétriques. En effet, la première, $V_{\Rightarrow}$, ne contient plus l'élement $A10 \Leftarrow B7$. Le seconde, $V_{\Leftarrow}$, ne contient plus $A5 \Rightarrow B2$ et $A8 \Rightarrow B4$.

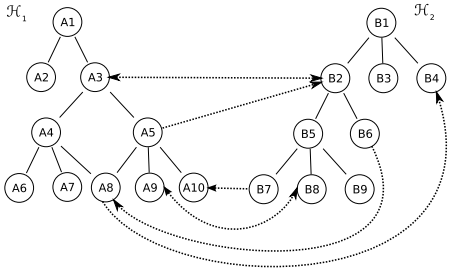


FIG. 2.8 – Alignement symétrique

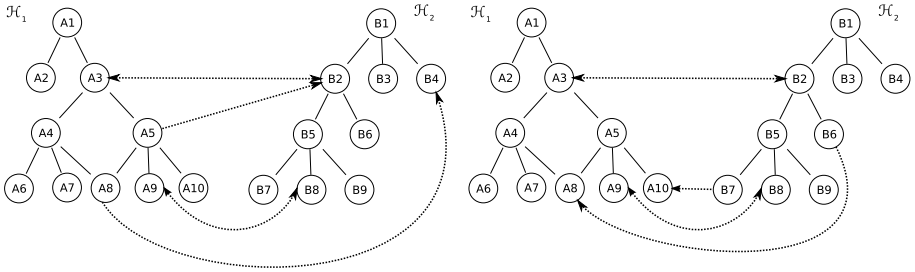


FIG. 2.9 – Composantes asymétriques $V_{\Rightarrow}$ et $V_{\Leftarrow}$

Un alignement peut être caractérisé par les cardinalités des couvertures minimales, $V^0_{\Rightarrow}$ et $V^0_{\Leftarrow}$, de ses deux composantes asymétriques $V_{\Rightarrow}$ et $V_{\Leftarrow}$. Dans le cas de $V^0_{\Rightarrow}$, nous considérerons le sens classique : un élément de C_1 peut être associé à un élément de C_2. Par contre, dans le cas de $V^0_{\Leftarrow}$, nous prendrons le sens inverse : un élément de C_2 peut être associé à un élément de C_1.

La cardinalité la moins restrictive est $0, n-0, m$. Elle signifie pour $V^0_{\Rightarrow}$ (resp. $V^0_{\Leftarrow}$) qu'une entité de C_1 (resp. C_2) peut être associée à 0, 1, ou plusieurs entités de C_2 (resp. C_1) et vice-versa. Nous distinguerons deux restrictions de la cardinalité. Premièrement, une composante asymétrique est fonctionnelle si la cardinalité de sa couverture minimale est $0, 1-0, n$. Cette cardinalité signifie, pour $V^0_{\Rightarrow}$ (resp. $V^0_{\Leftarrow}$), qu'une entité de C_1 (resp. C_2) est associée à, au plus, une entité de C_2 (resp. C_1). Deuxièmement, une composante asymétrique est injective si la cardinalité de sa couverture minimale est $0, n-0, 1$. Cette cardinalité signifie, pour $V^0_{\Rightarrow}$ (resp. $V^0_{\Leftarrow}$), qu'une entité de C_2 (resp. C_1) est associée à au plus 1 entité de C_1 (C_2). Une composante asymétrique peut être à la fois fonctionelle et injective, sa cardinalité sera alors notée $0, 1-0, 1$. La figure 2.10 montre les cardinalités de quatre alignements asymétriques (ou composantes asymétriques d'alignement).

Finalement, nous dirons qu'un alignement est fonctionnel (resp. injectif) si ses deux composantes asymétriques sont fonctionnelles (resp. injectives).

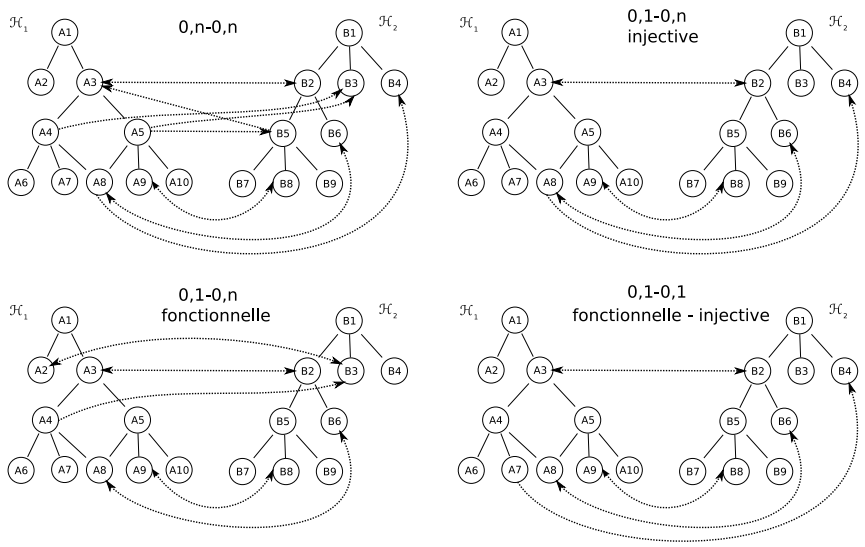


FIG. 2.10 – Cardinalités de composantes asymétriques minimales

2.3 Modèle de règles d'association entre hiérarchies

Dans le but d'extraire un alignement entre deux hiérarchies, nous proposons de transposer le paradigme des règles d'association introduit par [AIS93] aux hiérarchies. Les règles d'association étant un modèle d'extraction de connaissances à partir des données, la fouille de règles est effectuée sur des données extensionnelles. Le modèle de règles d'association entre hiérarchies est ainsi défini pour des hiérarchies contextualisées $\mathcal{H}_1 = (C_1, \leq, \mathcal{A}_1, O_1, \sigma_1)$ et $\mathcal{H}_2 = (C_2, \leq, \mathcal{A}_2, O_2, \sigma_2)$.

2.3.1 Contexte de fouille de données

La notion de règle d'association est définie sur un contexte de fouille constitué d'un ensemble d'individus E décrits par un ensemble de variables booléennes I au moyen d'une relation binaire $\delta \subseteq I \times E$. Généralement, ce contexte est une table (ou un ensemble de tables jointes) issue d'une base de données, décrite par un codage disjonctif complet. Dans le cadre de l'alignement de hiérarchies, notre contexte de fouille sera constitué des ensembles d'objets O_1 et O_2 qui seront décrits de par leur association aux entités C_1 et C_2 (considérées ainsi comme un ensemble de variables booléennes) par les relations binaires d'association σ_1 et σ_2.

Un contexte de fouilles de règles d'association entre hiérarchies est schématisé sur la figure 2.11. Il présente les deux hiérarchies $\mathcal{H}_1$ et $\mathcal{H}_2$ ainsi qu'un tableau binaire représentant les relations d'association des entités aux objets. Ce tableau est divisé verticalement en deux parties : la première représente

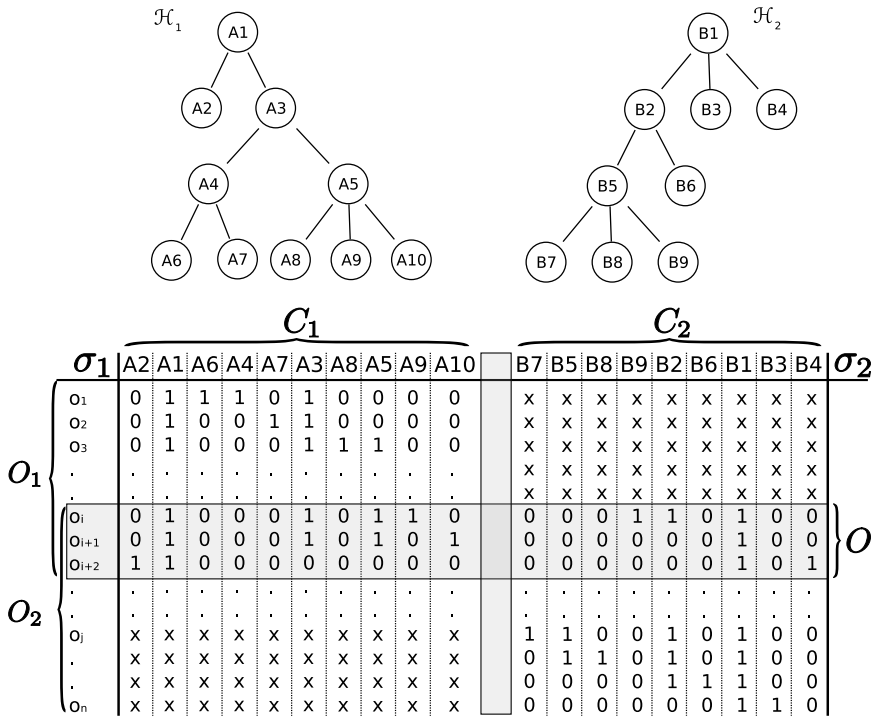


FIG. 2.11 – Contexte de fouille de règles entre hiérarchies

la relation d'association σ_1 et la deuxième, la relation d'association σ_2. Horizontalement, le tableau est séparé en trois parties : la première représente les objets de O_1 qui ne sont associés qu'aux entités C_1, la deuxième représente les objets communs aux hiérarchies $O = O_1 \cap O_2$, et la troisième représente les objets de O_2 qui ne sont associés qu'aux entités de $C2$. La valeur croisant un objet o_x ($o_x \in O_1 \cup O_2$) et une entité c_y ($c_y \in C_1 \cup C_2$) est 1 lorsque o_x est associé à c_y, 0 sinon. Une croix signifie que l'association entre o_x et c_y est indéfinie.

La fouille de règles ne peut se faire que sur un ensemble d'objets communs aux deux hiérarchies. Nous retiendrons ainsi l'ensemble $O = O_1 \cap O_2$, comme base pour l'évaluation des règles. Les relations d'association des entités aux objets σ_1 et σ_2 restreintes à cet ensemble seront alors notées σ.

2.3.2 Règles d'association entre entités

A partir de ce contexte de fouille, un alignement peut être déduit à partir des règles d'association entre les entités issues respectivement des deux hiérarchies. Une règle d'association $x \rightarrow y$ entre entités appartenant respectivement à C_1 et C_2, représente une tendance implicative de l'ensemble des objets associés à x, noté $\sigma(x)$ dans l'ensemble des objets associés à y, noté $\sigma(y)$.

Une règle d'association étant une tendance implicative, elle peut admettre quelques contre-exemples (c.-à-d. des objets de $\sigma(x)$ qui ne sont pas dans $\sigma(y)$), et l'inclusion stricte $\sigma(x) \subseteq \sigma(y)$, permettant d'induire l'implication $x \Rightarrow y$ (à partir de la propriété d'isomorphisme) n'est pas forcément vérifiée. Cette relaxe sur la contrainte d'inclusion est justifiée par le fait que les deux hiérarchies à aligner ne sont pas forcément décrites sur les mêmes données et qu'il est courant, dans la nature, d'observer quelques contre-exemples à une tendance générale sans que cette tendance ne puisse être contestée.

Une règle d'association peut s'interpréter : « Si un objet est associé à l'entité x alors il sera sûrement associé à l'entité y » ou encore « L'ensemble des objets associés à l'entité x a tendance à être inclus dans l'ensemble des objets associés à l'entité y ». Ainsi, partir d'une règle $x \rightarrow y$ jugée valide, nous pourrons déduire une implication de l'entité x vers l'entité y, notée $x \Rightarrow y$. La validité des règles sera jugée par une mesure d'intérêt permettant de vérifier la qualité inclusive de $\sigma(x)$ dans $\sigma(y)$ [CR06].

2.3.3 Différences par rapport à un contexte classique de fouille de règles

Le contexte de fouille défini ainsi que nos objectifs présentent, par rapport à une approche classique de fouille de règles d'association, les différences suivantes :

1. Les règles recherchées sont des règles binaires c.-à-d. qu'elles ne possèdent qu'une variable en prémisse et en conclusion.
2. Les prémisses et conclusions sont issues respectivement des ensembles de variables C_1 et C_2 qui sont disjoints.
3. Les ensembles de variables C_1 et C_2 sont chacun munis d'une relation d'ordre partiel.
4. Les ensembles d'objets O_1 et O_2 ne sont pas égaux. L'évaluation des règles ne peut être faite que sur les objets partagés $O_1 \cap O_2$.

Les deux premières différences constituent un avantage par rapport aux contextes classiques de fouilles de règles. En effet, traditionnellement, un algorithme d'extraction des règles d'association doit générer l'ensemble des combinaisons fréquentes d'attributs (appelées également itemsets fréquents), puis envisager toutes les règles possibles entre ensembles d'attributs fréquents. La taille de l'espace de recherche pour l'extraction des combinaisons fréquentes d'attributs croît de manière exponentielle avec le nombre p d'attributs (il y a potentiellement $2^p - 1$ combinaisons fréquentes d'attributs). Dans notre cas, nous aurons, au pire, seulement $|C_1| \times |C_2|$ évaluations de règles à effectuer.

La deuxième différence va engendrer le problème de redondance dans l'ensemble des règles extraites (voir section 2.2.3). La présence de relation d'ordre (ou taxonomie) entre variables a été posée par [SA95]. Les auteurs ont proposé une solution visant à extraire les règles les plus générales, c.-à-d. celles qui possèdent les variables les plus générales possibles en terme de support. Dans notre cas, ces règles ne nous intéressent pas. En effet, nous voulons en accord avec le principe de redondance proposé par l'équation 2.12 des règles ayant une

prémisse générale et une conclusion spécifique [Leh00] et [PTB+05]. Nous utiliserons, pour cela, une approche combinant les notions de règles généralisées, pour le prise en compte des taxonomies sur les items, et celle de la redondance de [Leh00] et [PTB+05]. Cette approche a été exposée dans la section 1.3.

La dernière différence est la plus problématique. En effet, comme nous inférons l'alignement à partir des données extensionnelles, les deux hiérarchies doivent être comparées sur une base commune. Cependant, l'ensemble des données extensionnelles partagées par les deux structures, c.-à-d. $O = O_1 \cap O_2$, est souvent trop petit (ou même vide) pour que des inférences statistiquement valides puissent être faites. Ce problème de données non partagées est commun à toutes les méthodes d'alignement extensionnelles. Certaines méthodes proposent ainsi des méthodes de prétraitement. Ces méthodes permettent, soit de classer les objets de chaque hiérarchie dans l'autre à l'aide de techniques d'apprentissage supervisées, soit de redéfinir l'extension des hiérarchies sur des descripteurs issus des objets.

Conclusion

Nous avons exposé dans ce chapitre nos modèles de hiérarchie et d'alignement. Dans la section traitant des hiérarchies, nous avons distingué la hiérarchie hors-contexte qui est constituée uniquement d'un ensemble d'entités organisées par une relation d'ordre partiel. Nous avons ensuite présenté le modèle des hiérarchies contextualisées dans lequel chaque entité est associée à un ensemble d'objets. Ce type de hiérarchies est largement utilisé en informatique. Elles servent à représenter des catalogues de boutiques en ligne, des répertoires de sites web, thésaurus auxquels on a indexé des documents, la structure d'un document XML et même de manière générale, des systèmes de fichiers.

Dans la deuxième section, nous avons présenté notre modèle d'alignement. Un alignement peut être défini comme un ensemble d'éléments de correspondance entre entités issues de deux hiérarchies. Nous insistons sur le fait qu'une correspondance peut être non seulement une équivalence mais également une implication. Nous avons introduit des règles permettant de déduire, à partir de l'alignement et des hiérarchies, de nouveaux éléments de correspondance. Grâce à ces règles, nous avons ensuite défini les notions de fermeture et couverture minimale d'un alignement. Ces dernières notions nous ont permis de formaliser la redondance dans un alignement et sa consistance. Finalement, nous avons étudié la symétrie et la cardinalité d'un alignement.

Dans la dernière section, nous avons présenté l'adaptation du paradigme des règles d'association à l'alignement de hiérarchies. C'est sur ce modèle d'alignement extensionnel de hiérarchies que notre méthode d'alignement, présentée dans le chapitre 4, est fondée.

Les méthodes d'alignement de hiérarchies 3

Sommaire

Introduction

D'une manière générale, l'alignement vise à identifier un ensemble d'appariements entre éléments issus de deux représentations structurées distinctes.

La problématique de la recherche d'alignement entre deux structures a été posée dans de nombreux domaines. Dans le domaine des bases de données, par exemple, la recherche d'alignement entre schémas est étudiée depuis le début des années 80 pour des objectifs d'intégration et d'interopérabilité [BLN86]. Un autre exemple est celui de la bioinformatique où la recherche d'un alignement de séquences vise à identifier des portions communes ou similaires entre deux séquences d'ADN, d'ARN ou de protéines [NW70]. Des recherches ont été également initiées dans le domaine des graphes afin de détecter des isomorphismes partiels entre deux graphes.

De nombreuses méthodes d'alignement dédiées aux schémas (bases de données, objets, xml) et aux ontologies ont vu le jour cette dernière décennie. Ces méthodes ont été étudiées dans plusieurs états de l'art ([SE05], [KS03], [RB01]) qui les ont classées en fonction des techniques, des types d'information, des critères qu'elles utilisent, et de la façon dont elles les exploitent.

Dans ce chapitre, nous étudions les méthodes d'alignement en utilisant tout d'abord un point vue global. Pour cela, nous distinguons leurs caractéristiques externes (entrées et sorties) et la manière dont elles composent les différentes techniques qu'elles utilisent. Ensuite, nous focalisons sur les techniques permettant la comparaison des entités. Ces techniques peuvent être basées soit sur la description intensionnelle (le schéma) des entités, soit sur les données extensionnelles (les instances) qui leur sont associées. Pour la première famille de techniques, nous nous appuyerons particulièrement sur les critères de classification proposés dans [SE05] et [RB01]. Finalement, nous proposons une classification d'une vingtaine de méthodes d'alignement, synthétisant les caractéristiques externes, la composition et les techniques d'alignement qu'elles utilisent.

3.1 Définitions et notations

Cette section a pour premier objectif d'introduire la définition que nous donnons d'une méthode d'alignement de hiérarchies. Dans un second temps, nous faisons quelques rappels sur les définitions de distances et similarités qui sont des mesures largement utilisées par les méthodes d'alignement.

3.1.1 Méthode d'alignement

Etant donné deux hiérarchies $\mathcal{H}_1 = (C_1, \leq, \mathcal{A}_1, O_1, \sigma_1)$ et $\mathcal{H}_2 = (C_2, \leq , \mathcal{A}_2, O_2, \sigma_2)$, une procédure d'alignement permet de comparer les entités $x \in C_1$ et $y \in C_2$ afin d'extraire un alignement A entre les deux ensembles C_1 et C_2.

On trouve différentes définitions de l'alignement. Dans un contexte d'alignement de schémas de bases de données, E. Rahm et P.A. Bernstein [RB01] définissent une méthode d'alignement comme une fonction qui prend en entrée deux schémas $S1$ et $S2$ (eux-mêmes définis comme un ensemble d'éléments connectés par une structure quelconque) et qui retourne un appariement entre l'ensemble des élements de $S1$ vers celui de $S2$. Chaque élément de l'appariement (retourné par la méthode d'alignement) exprime une relation entre un élément de $S1$ et un élément de $S2$.

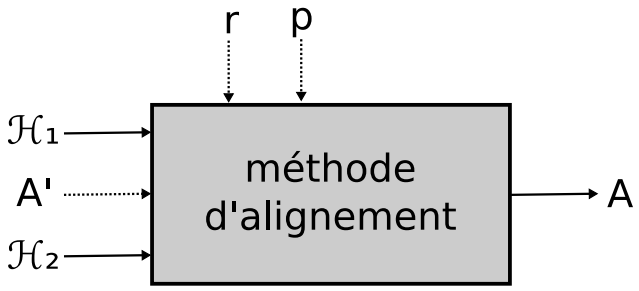


FIG. 3.1 – Entrées/sorties d'une méthode d'alignement

Une définition plus complète a été proposée par [SE05] dans le contexte d'alignement d'ontologies. Elle rajoute notamment la notion de paramètres que peut en compte la méthode. Ces paramètres peuvent être un alignement d'entrée (qui sera ainsi enrichi par la méthode), les ensembles de valeurs de seuil et de pondérations utilisées par les algorithmes, et également les ressources externes (comme par les thésaurus) sur lesquelles s'appuie la méthode.

En reprenant la dernière définition, nous définissons une méthode d'alignement de hiérarchies, schématisée figure 3.1, de la manière suivante :

Définition 3.1 *Une procédure d'alignement est une fonction M prenant en entrée deux hiérarchies $\mathcal{H}_1$ et $\mathcal{H}_2$ et retournant une structure A appelée alignement. Cette procédure d'alignement peut également être configurée par un ensemble de paramètres (seuils de sélection, pondérations, etc.) p, avoir recours à un ensemble de ressources externes (ressources terminologiques, dictionnaires, etc.) r et utiliser un alignement préliminaire A'.*

Nous pouvons noter $M(\mathcal{H}_1, \mathcal{H}_2, p, r, A') = A$, mais lorsque le contexte est supposé fixe, nous ometterons les paramètres et utiliserons la notation simplifiée $M(\mathcal{H}_1, \mathcal{H}_2) = A$.

3.1.2 Comparaison d'individus : distances et similarités

Les mesures de distance ou de similarité permettent d'évaluer l'éloignement, ou la ressemblance entre deux éléments (ou individus) issus d'un même ensemble I. Nous considérons qu'un individu a est décrit par un ensemble de caractéristiques, noté A.

Distance

Une distance est une application de $I \times I$ dans $\mathbb{R}^+$ telle que :

$$\begin{aligned} d(a,b) &= d(b,a) \\ d(a,b) &\geq 0 \\ d(a,b) = 0 &\Leftrightarrow a = b \\ d(a,b) &\leq d(a,c) + d(c,b) \end{aligned}$$

On parlera de dissimilarité si la mesure vérifie seulement :

$$\begin{aligned} d(a,b) &= d(b,a) \\ d(a,b) &\geq 0 \\ d(a,a) &= 0 \end{aligned}$$

Un exemple de distance utilisée par les méthodes d'alignement est la distance de Hamming [Ham50]. Cette distance représente la différence symétrique des ensembles de caractéristiques associés à deux individus a et b :

$$d_{Hamming}(a,b) = |A \Delta B| = |A - B| + |B - A|$$

Similarité

Une mesure de similarité évalue la ressemblance entre deux individus a et b. C'est une application de $I \times I$ dans $[0;1]$ telle que :

$$s(a,b) = s(b,a) \tag{3.1}$$
$$s(a,b) = 0 \Rightarrow A \cap B = \emptyset \tag{3.2}$$
$$s(a,b) = 1 \Leftrightarrow a = b \tag{3.3}$$

De nombreuses mesures de similarité ont été proposées dans la littérature. Nous en donnons ici quelques définitions de mesures que nous utiliserons par la suite.

La mesure de Jaccard [Jac01] :

$$s_{Jaccard}(a,b) = \frac{|A \cap B|}{|A \ \cup B|}$$

La mesure de Dice [Dic45] :

$$s_{Dice}(a,b) = \frac{2|A \cap B|}{|A| + |B|}$$

La mesure d'Occhai [Och57] :

$$s_{Occhai}(a,b) = \frac{|A \cap B|}{\sqrt{|A| + |B|}}$$

Les mesures d'Occhai et de Dice peuvent être généralisées en une famille de mesures proposée par [CK96]. Ces mesures sont définies comme le rapport du nombre de caractères partagés par a et b sur une moyenne de Cauchy entre le nombre de caractères de a et de b :

$$s_\alpha(a,b) = \frac{|A \cap B|}{m_\alpha(|A|,|B|)}$$

où m_α est définie par :

$$m_\alpha(|A|,|B|) = \sqrt[\alpha]{\frac{|A|^\alpha + |B|^\alpha}{2}}$$

Dans le cas où $\alpha \to 0$, on obtient la mesure d'Occhai et quand $\alpha = 1$, on a celle de Dice. Si $\alpha \to -\infty$, alors $m_\alpha(|A|,|B|) = min(|A|,|B|)$ et si $\alpha \to +\infty$, alors $m_\alpha(|A|,|B|) = max(|A|,|B|)$.

On peut également noter que les mesures de Jaccard et Occhai appliquées au modèle vectoriel sont respectivement appelées coefficients de Taminoto et du cosinus.

3.2 Caractéristiques externes d'une méthode d'alignement

Sans rentrer dans le détail du fonctionnement d'une méthode d'alignement, nous pouvons dans un premier temps distinguer différentes familles de méthodes par rapport à leurs caractéristiques externes. Ces caractéristiques externes concernent les entrées, les sorties et les différents paramètres des méthodes d'alignement.

3.2.1 Données d'entrées

On peut d'abord s'intéresser à leurs données d'entrée [SE05]. En effet, il existe de nombreux formalismes de représentation de données hiérarchiques disponibles. Une méthode d'alignement est généralement conçue pour fonctionner avec certains types de formalismes. Une méthode peut ainsi considérer des schémas de base de données (de type relationnel, XML, ou objet), des ontologies (décrites en OWL, RDFS, ou autre language), ou même de simples structures arborescentes (structures des répertoires Web, structures des répertoires d'un système de fichiers). On peut également distinguer les approches en fonction de la nécessité ou non de leur fournir des hiérarchies peuplées. En effet, certaines méthodes s'appuient sur les données extensionnelles des hiérarchies, c.-à-d. les instances d'une base de données objet ou d'une ontologie, les n-uplets d'une base relationnelle, le contenu des documents XML ou encore les fichiers d'un système de fichiers, etc.

3.2.2 Données de sorties

Dans un second temps, on peut également remarquer des différences entre méthodes quant à l'alignement qu'elles produisent. Une méthode peut assigner

ou non une valeur de qualité aux éléments de correspondances trouvés. En effet, il existe, d'une part, les approches symboliques, et d'autre part, les approches basées sur des similarités. Les premières vont seulement permettre de déterminer si deux entités sont en correspondance ou non, tandis que les deuxièmes vont également quantifier la confiance que l'on doit accorder aux éléments de correspondance qu'elles retournent.

Les méthodes se distinguent également par rapport aux types de relations qu'elles peuvent trouver en deux entités. En effet, la grande majorité des méthodes s'intéressent seulement à la relation d'équivalence ($\Leftrightarrow$[1]), cependant certaines méthodes sont plus expressives et distinguent, notamment, les relations d'implication ($\Rightarrow$[2] et $\Leftarrow$).

Une procédure d'alignement peut être symétrique ou asymétrique. Elle est symétrique si l'ordre dans lequel les hiérarchies sont considérées n'a pas d'influence sur l'alignement produit. En d'autres termes, si $M(\mathcal{H}_1, \mathcal{H}_2) = M(\mathcal{H}_2, \mathcal{H}_1)$ alors M est symétrique sinon (si $M(\mathcal{H}_1, \mathcal{H}_2) \neq M(\mathcal{H}_2, \mathcal{H}_1)$) M est asymétrique. Si l'on considère la relation d'implication, une méthode asymétrique produira des alignements asymétriques tandis qu'une méthode symétrique produira des alignements symétriques. La notion de symétricité des alignements a été introduite section 2.2.5.

Une méthode d'alignement se distingue également par la cardinalité des alignements qu'elle produit. Dans le cadre général, nous avons vu qu'un alignement entre deux hiérarchies $\mathcal{H}_1$ et $\mathcal{H}_2$ possède deux cardinalités : l'une dans le sens $\mathcal{H}_1$ vers $\mathcal{H}_2$ et l'autre dans le sens $\mathcal{H}_2$ vers $\mathcal{H}_1$. Cette distinction n'est nécessaire que pour les méthodes prenant en compte des relations directionnelles telles que l'implication. Cependant, comme la grande majorité des méthodes s'intéressent uniquement à l'équivalence, cette distinction n'est pas faite dans la littérature. Principalement, une méthode d'alignement peut produire des alignements :

- fonctionnels (de cardinalité $0, 1 - 0, n$),
- fonctionnels et injectifs (de cardinalité $0, 1 - 0, 1$),
- de cardinalité $0, n - 0, m$.

La plupart des méthodes d'alignement produisent des alignements au moins fonctionnels. En effet, elles mettent en correspondance chaque élément d'une hiérarchie source à, au maximum, un élément de la hiérarchie cible. Une méthode qui produit des alignements fonctionnels mais pas injectifs sera asymétrique. Par contre, une méthode produisant des alignements fonctionnels et injectifs n'est pas forcément symétrique.

[RB01] distingue la cardinalité globale, c.-à-d. celle dont on a discuté précédemment, de la cardinalité locale. Cette cardinalité locale intervient lorsqu'une méthode permet également l'alignement des constituants d'une entité (comme par exemple, les attributs d'une table relationnelle, ou d'une classe objet). Dans le cadre de nos travaux, nous ne nous intéressons pas à l'alignement des constituants d'entités et ne prenons pas en compte ce niveau de cardinalité.

[1] notée également $=$

[2] notées également, en fonction du contexte, $\sqsubseteq$, $\leq$

3.3 Composition interne des méthodes

Une méthode d'alignement réalise, la plupart du temps, une combinaison de plusieurs approches d'appariements. En effet, afin d'obtenir des résultats convenables une méthode doit utiliser plusieurs critères, s'appuyer sur diverses techniques et utiliser différents niveaux d'information sur la description des entités à apparier.

Du point de vue de [RB01], il existe deux manières de combiner différentes approches au sein de la méthode globale :

- L'hybridation permettant de combiner, en un seul algorithme, plusieurs approches basées sur différents critères et types d'information ;
- La composition permettant de combiner les résultats produits par plusieurs algorithmes exécutés de manière indépendante. Ces algorithmes d'alignement peuvent être basés sur des approches individuelles (utilisant un seul critère et source d'information) ou des approches hybrides.

Les méthodes hybrides ont l'avantage d'obtenir de meilleures performances que l'exécution de plusieurs algorithmes individuels séparément. En effet, la prise en compte de plusieurs critères simultanés permet d'écarter directement des relations n'en satisfaisant qu'un seul et ainsi d'éviter de parcourir en totalité et/ou plusieurs fois les structures à aligner. Par exemple, la prise en compte de la relation d'ordre par un parcours préfixé (du haut vers le bas) des hiérarchies permet, lorsqu'une entité est incompatible avec une autre, d'éviter l'évaluation, inutile, d'appariements entre leurs subsumées. L'approche hybride est aussi intéressante pour prendre en compte des relations qui n'auraient pas été considérées en utilisant un seul critère.

Lorsque les critères étudiés sont indépendants, il est alors plus intéressant d'utiliser une approche composite qui permet plus de flexibilité. Par exemple, les critères utilisés pour aligner des ontologies décrites en OWL et des thésaurus SKOS ne seront pas les mêmes. Les méthodes composites ont l'avantage d'être modulaires et ainsi d'être adaptables plus facilement à différentes représentations de structures d'entrée. Même si dans la plupart des cas, la composition est intrinsèquement liée aux méthodes[3], les auteurs de la méthode COMA (et ses dérivées) [DR02], ont exploité le coté modulaire de cette approche afin de permettre le choix à l'utilisateur de combiner différents algorithmes à sa guise.

Cependant, les méthodes composites nécessitent de fusionner les résultats produits par les différents algorithmes. La stratégie de fusion des résultats dépend des types d'algorithmes individuels ou hybrides utilisés et surtout de la manière dont ils sont combinés. Nous distinguons deux approches principales de composition d'algorithmes :

- **la composition parallèle** où les résultats des algorithmes individuels sont obtenus de manière indépendante, puis agrégés pour former l'alignement final ;
- **la composition linéaire ou séquentielle** où les résultats produits par

[3] [RB01] ont tendance à regrouper ce type de méthodes (c.-à-d. les méthodes dont la composition est réalisée en « dur ») dans la catégorie des méthodes hybrides. Dans notre notre cas, nous dirons qu'une méthode est composite dès lors que les critères sont étudiés de manière indépendante

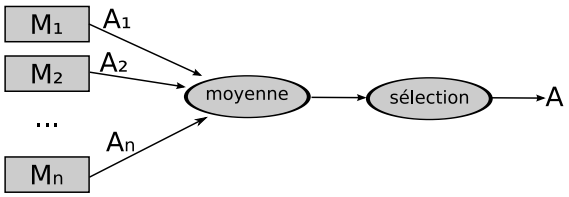


FIG. 3.2 – Combinaison statistique

un algorithme servent d'entrée à un suivant et ainsi de suite. La production de l'alignement final est ainsi successivement raffinée.

Au sein d'une même méthode, les différentes approches de composition peuvent être utilisées.

3.3.1 Composition parallèle

La composition parallèle permet de combiner les résultats produits par un ensemble d'algorithmes individuels ou hybrides exécutés de manière totalement indépendante [RB01]. En fonction du type d'algorithme utilisé, nous avons recensé deux approches de combinaison des résultats.

Le premier cas concerne les algorithmes ne permettant pas la sélection d'éléments de correspondance. En effet, certains algorithmes sont seulement des fonctions calculant une valeur de qualité pour chaque appariement possible (c.-à.-d. pour chaque couple d'entités et pour chaque relation étudiée). Dans ce cas, la combinaison des résultats consiste à (1) agréger les fonctions de qualité produites par chaque algorithme en une fonction globale, puis à (2) sélectionner un sous-ensemble d'éléments de correspondances pertinentes. Nous appelons cette première approche d'agrégation des résultats, combinaison statistique.

Le deuxième cas concerne les algorithmes d'alignement à proprement parler, qui permettent de sélectionner un ensemble d'éléments de correspondance pertinents et de fournir, éventuellement, une fonction de qualité. Dans ce cas, la combinaison des résultats consiste à (1) réaliser une opération ensembliste (union ou intersection) sur les ensembles de correspondances produits par chaque algorithme et à (2) agréger les fonctions de qualité lorsqu'elles sont définies sur des éléments de correspondances produits par plusieurs algorithmes. Nous appelons cette deuxième approche, combinaison ensembliste des résultats intermédiaires.

Combinaison statistique

Dans une approche par combinaison statistique des résultats (schématisée figure 3.2), chaque algorithme M_i produit un alignement $A_i(C_1 \times C_2 \times R, q_i)$ où $R \subseteq \{\Rightarrow, \Leftarrow, \Leftrightarrow\}$.

La première étape consiste à agréger les fonctions de qualités q_i en une fonction de qualité globale q. Cette fonction globale est définie pour chaque

triplet $a = (x, y, r) \in C_1 \times C_2 \times R$ par une moyenne des fonctions q_i :

$$q(a) = \sqrt[m]{\frac{1}{n}\sum_{i=1}^{n} q_i(a)^m}$$

où m peut être égale à :
- $m \to -\infty$: minimum de la série.
- $m = -1$: moyenne harmonique,
- $m \to 0$: moyenne géométrique (3.3.1),
- $m = 1$: moyenne arithmétique,
- $m = 2$: moyenne quadratique.
- $m \to +\infty$: maximum de la série.

Cette moyenne peut être également pondérée par des confiances p_i associées à chaque algorithme. Dans le cas d'une moyenne arithmétique ($m = 1$), elle est définie ainsi :

$$q(a) = \frac{\sum_{i=1}^{n} p_i.q_i(a)}{\sum_{i=1}^{n} p_i}$$

Si les valeurs de confiance retournées par les algorithmes sont vues comme des probabilités indépendantes les unes des autres alors on a recours à une moyenne géométrique définie ainsi :

$$q(a) = \sqrt[m]{\prod_{i=1}^{n} q_i(a)^m}$$

De la même manière que pour la moyenne arithmétique, une moyenne géométrique peut être pondérée par des confiances p_i associées à chaque algorithme :

$$q(a) = \left(\prod_{i=1}^{n} q_i(a)^{p_i}\right)^{\frac{1}{\sum_{i=1}^{n} p_i}}$$

Une fois que les fonctions de qualités sont fusionnées, la deuxième étape consiste à sélectionner un ensemble d'éléments de correspondance pertinents afin de produire un alignement $A = (V, q)$. Il existe plusieurs stratégies de sélection des éléments de correspondance :
- seuillage sur les valeurs de qualité,
- maximisation locale de l'appariement,
- maximisation globale de l'appariement.

La sélection de l'ensemble V peut être réalisée premièrement par seuillage. Dans ce cas, un élément a sera sélectionné si $q(a)$ est supérieure à un seuil minimum q_0. L'ensemble des éléments de correspondance V est alors défini par :

$$V = \{a \in C_1 \times C_2 \times R | q(a) \geq q_0\}$$

En suivant cette stratégie, l'alignement produit n'aura aucune contrainte sur sa cardinalité. En effet, une entité d'une des hiérarchies pourra être en correspondance avec plusieurs entités de l'autre hiérarchie, du moment que les éléments de correspondance dans lesquels elle figure ont une valeur de qualité supérieure au seuil choisi.

La seconde stratégie de sélection est basée sur la maximisation locale de l'appariement. La maximisation locale consiste à choisir pour chaque entité de la hiérarchie source, l'élément (ou les éléments) de correspondance maximisant la valeur de qualité.

$$V = \{a = (x, y, \mathcal{R}) \in C_1 \times C_2 \times R | \forall a' = (x, y', \mathcal{R}') \in C_1 \times C_2 \times R, \\ q(a') < q(a) \wedge a \neq a'\} \quad (3.4)$$

Cette stratégie de sélection rend la méthode d'alignement asymétrique. En effet, un élément de correspondance $a = (x, y, \mathcal{R})$ ayant la meilleure valeur de qualité pour x n'est pas forcément celui qui a la meilleure valeur de qualité pour y. L'ordre dans lequel sont considérées les hiérarchies devient ainsi important. De plus, cette stratégie aura tendance à produire des alignements fonctionnels (de cardinalité $0, 1-0, n$) dans le sens $\mathcal{H}_1$ vers $\mathcal{H}_2$ s'il existe pour chaque entité de la hiérarchie source un seul élément de correspondance maximisant la valeur de qualité. Cette stratégie de sélection des correspondances est celle qui est la plus utilisée par les méthodes d'alignement [RB01]. Il existe également des alternatives basées sur la minimisation locale du risque [TLL+06].

Une autre stratégie s'appuie sur la maximisation globale de l'appariement. Elle consiste à sélectionner parmi un ensemble d'alignements répondant à un certain critère de cardinalité, celui qui maximise la somme des valeurs de qualité. Ce principe a été exposé dans [Val99] pour définir une distance d'alignement entre deux ensembles et utilisé par [EBB+04] au sein d'une méthode d'alignement d'ontologies ne considérant que l'équivalence. Ce principe est plus difficile à mettre en place dans le cadre général d'alignement de hiérarchies où chaque ensemble est muni d'une relation d'ordre et où un alignement prend en compte différents types de relations.

Soit M l'ensemble des ensembles de correspondance possibles. Si l'on fait abstraction des différents types de relations R, M sera constitué de triplets $(x, y, \mathcal{R})$ où x et y n'apparaissent, respectivement, qu'une seule fois dans M. Les types de relations peuvent être considérées de manière indépendante. Dans ce cas, M sera constitué de triplets $(x, y, \mathcal{R})$ où x et y, n'apparaissent, respectivement, qu'une seule fois dans M pour un type de relation $\mathcal{R}$ donné. Finalement, on peut également considérer la sémantique des relations et associer à chaque type de relation un modèle de cardinalité. Par exemple, on peut associer la cardinalité $0, 1-0, 1$ pour la relation $\Leftrightarrow$, $0, 1-0, n$ (respectivement $0, n-0, 1$) pour la relation $\Rightarrow$ (respectivement $\Leftarrow$).

A partir d'un ensemble M, l'ensemble des éléments de correspondance sélectionnés sera défini par :

$$V = argmax_{m \in M} \frac{\sum_{a_i \in m} q(a_i)}{min(|C_1|, |C_2|)}$$

[Val99] a utilisé ce principe pour définir une distance globale entre deux ensembles mais en normalisant par $max(|C_1|, |C_2|)$.

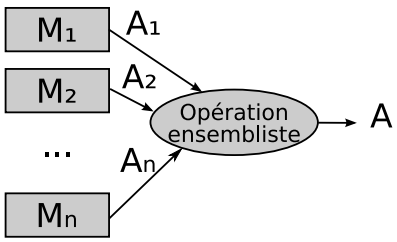


FIG. 3.3 – Combinaison ensembliste

Combinaison ensembliste

La combinaison ensembliste des résultats est utilisée lorsque chaque algorithme M_i produit un alignement $A_i(V_i, q_i)$ où $V_i \subseteq C_1 \times C_2 \times R$. Dans ce cas, la fusion des résultats intermédiaires consiste à réaliser une opération ensembliste $\bigcup$ ou $\bigcap$ sur les ensembles de correspondances V_i. En fonction de l'opération utilisée, l'ensemble final d'éléments de correspondances V sera défini de l'une des deux façons suivantes :

$$V = \bigcup_{i=1}^{n} V_i$$

$$V = \bigcap_{i=1}^{n} V_i$$

La combinaison ensembliste avec union est utilisée lorsque par exemple chaque algorithme est spécialisé dans un type de relation ou sur certaines parties des hiérarchies. La combinaison avec intersection peut être quant à elle pertinente lorsque l'on désire qu'une relation ne soit validée que si elle respecte tous les critères considérés par les algorithmes. Ce dernier type de combinaison est particulièrement exigeant.

Une combinaison ensembliste peut être également utile lorsque l'on veut symétriser les résultats produits par une méthode M non symétrique. En effet, si l'on désire passer à partir d'une méthode de cardinalité $0,1-0,n$ à un alignement symétrique de cardinalité $0,n-0,m$ alors on peut réaliser une combinaison ensembliste par union, $M(\mathcal{H}_1, \mathcal{H}_2) \cup M(\mathcal{H}_2, \mathcal{H}_1)$. Si l'on veut passer à un alignement symétrique de cardinalité $0,1-0,1$ alors on réalisera l'intersection $M(\mathcal{H}_1, \mathcal{H}_2) \cap M(\mathcal{H}_2, \mathcal{H}_1)$.

Lorsqu'un élément de correspondance sélectionné est présent dans plusieurs ensembles V_i, la fonction de qualité globale q sera déduite à partir des fonctions q_i en utilisant un des moyens d'agrégation des fonctions de qualité présentés dans la section précédente. La fonction q sera alors soit une moyenne, soit le maximum ou soit le minimum des valeurs des fonctions q_i.

3.3.2 Composition linéaire

Dans une composition linéaire, les algorithmes sont exécutés successivement et le résultat d'un algorithme sert d'entrée à un suivant et ainsi de suite jusqu'à obtenir l'alignement final. La composition linéaire est utilisée par deux types d'algorithmes : les algorithmes de découverte d'alignement, et les algorithmes de filtre d'alignement.

La composition linéaire dans les méthodes d'alignement est utilisée lorsque qu'un algorithme de découverte de relations se base sur un alignement préalable afin de l'étendre et de l'affiner. Ce type de composition est souvent utilisé lorsque l'algorithme suit une stratégie d'alignement structurelle (voir section 3.4.2). Cupid [MBR01], Similarity Flooding [MGMR02] ou encore GLUE [DMDH02] sont des exemples de méthode d'alignement utilisant ce type de composition linéaire.

L'autre utilisation de la composition linéaire concerne les filtres d'alignement. Un algorithme de filtre d'alignement permet de réduire le nombre d'éléments d'un alignement en fonction de multiples critères. Un filtre d'alignement produit à partir d'un alignement $A' = (V', q)$ donné en entrée, un alignement filtré $A = (V, q)$ tel que $V \subseteq V'$. Les critères permettant de filtrer un alignement peuvent porter sur la cardinalité, la consistance, la redondance, etc.

Un filtre peut permettre de réduire la cardinalité d'un alignement. Par exemple, il est possible de passer d'un alignement de cardinalité $0, n-0, m$ (dans le sens $\mathcal{H}_1$ vers $\mathcal{H}_2$) à un alignement de cardinalité $0, 1-0, m$ en sélectionnant seulement pour chaque entité x, l'élément $(x, y, \mathcal{R})$ de V' qui maximise la fonction de qualité q. Ce principe est l'un des plus utilisé par les méthodes d'alignement [RB01]. L'ensemble des éléments de correspondances V sera ainsi défini par :

$$V = \{a = (x, y, \mathcal{R}) \in V' \mid \forall a' = (x, y', \mathcal{R}') \in V', q(a') < q(a)\}$$

Les filtres portant sur la consistance vont vérifier pour chaque élément d'un alignement s'il n'existe pas un autre élément avec lequel il est en contradiction. La notion d'éléments de correspondance en contraction est introduite dans la section 2.2.4. Lorsqu'une correspondance est en contradiction avec une autre (ou un ensemble d'autres), plusieurs stratégies peuvent être utilisées pour résoudre le problème. La première stratégie consiste à éliminer une correspondance a en contradiction avec une correspondance a', si la valeur de confiance associée à a est plus petite que celle associée à a'. Une deuxième stratégie consiste à compter pour chaque correspondance a, le nombre de correspondances avec lesquelles elle est en conflit. Les correspondances les plus conflictuelles seront ainsi itérativement supprimées jusqu'à ce que le nombre de conflits soit nul ou au-dessous d'un seuil fixé.

Les filtres de réduction de redondance (ou de couverture minimale) vont permettre de vérifier la minimalité d'un alignement. Plus précisément, ce type d'algorithme vise à éliminer les éléments de correspondance pouvant être déduits à partir d'autres. La redondance dans un alignement n'est possible que si la méthode permet de découvrir des implications entre entités. L'ensemble des éléments de correspondance V obtenu par élimination de la redondance à partir de l'ensemble V' est égal à sa couverture minimale V'^O (voir section 2.2.3).

3.4 Les techniques d'alignement intentionnelles

La description intentionnelle d'une hiérarchie $\mathcal{H} = (C, \leq, \mathcal{A}, O, \sigma)$ est constituée de sa partie hors-contexte. Cette description comprend les entités C, les fonctions d'annotations $\mathcal{A}$, ainsi que la relation d'ordre $\leq$. En principe, l'ensemble des fonctions d'annotation contient au moins la fonction assignant un identifiant (ou nom) à chaque entité d'une hiérarchie. Selon le type de schéma (catalogue web, annuaire, thésaurus, ontologie, schéma objet ou de base de données) et également le langage de représentation utilisé, les entités peuvent être décrites de manière plus ou moins expressive. En effet, certains langages vont permettre de définir des fonctions d'annotation permettant d'associer, également, aux entités des labels (parfois en plusieurs langues) et des commentaires.

Sur des ontologies, ou des schémas de bases de données ou objet, la description intensionnelle d'une hiérarchie est également pourvue d'un ensemble de relations. Ces relations peuvent être des relations transversales inter-entités ou encore des propriétés (ou attributs) définies sur des types de données simples (voir section 2.1).

Parmi les méthodes s'appuyant sur les descriptions intensionnelles, [EBB+04] et [SE05] ont distingué les méthodes terminologiques basées sur les descriptions textuelles des approches structurelles qui peuvent s'appuyer sur la relation d'ordre, mais également sur les attributs et les relations transversales entre entités.

3.4.1 Techniques terminologiques

Les méthodes terminologiques s'appuient sur la comparaison des chaînes de caractères afin d'en déduire une similarité (ou dissimilarité) ou une relation d'hyperonomie/hyponymie. Le moyen le plus simple pour comparer deux chaînes de caractères consiste à comparer leur structure : plus elles partageront de caractères ou mots en commun, plus elles seront similaires. Les méthodes s'appuyant sur cette première approche sont appelées méthodes terminologiques syntaxiques. D'un autre coté, les méthodes terminologiques linguistiques comparent deux chaînes de caractères en ayant recours à une base de données lexicale sous forme de réseau sémantique. A partir de ce type de ressources terminologiques, ces méthodes permettent de calculer une similarité ou de déduire la relation qu'entretiennent deux termes.

Techniques terminologiques syntaxiques

En fonction du modèle de similarité utilisé, une chaîne de caractères A est modélisée de différentes façons :

- Une séquence de caractères notée $A = (a_1; ...; a_n)$ où a_x sont des lettres (ou symboles).
- Une séquence de sous-chaînes de caractères (appelées mots) séparées par des délimiteurs (espaces, tirets bas, ponctuation). Dans ce cas, une chaîne de caractères sera représentée par une séquence de mots notée

$A = (A_1; ...; A_m)$ où A_x est une chaîne de caractères.
Le moyen le plus simpliste de comparaison de deux chaînes de caractères est de vérifier leur identité. La similarité sera alors binaire, c.-à-d. égale à 1 lorsque $A = B$ et 0 lorsque $A \neq B$.

Un modèle simple pour comparer deux séquences de caractères est d'ignorer l'ordre d'apparition des lettres dans la séquence. Dans ce cas, la comparaison de deux chaînes a et b, représentées respectivement par les ensembles de caractères A et B, consiste à utiliser une mesure de similarité ou une distance ensembliste. Des mesures couramment utilisées sont la distance de Hamming ou la similarité de Jaccard (voir définitions de la section 3.1.2).

Par contre, si l'on prend en compte l'ordre d'apparition des lettres, une famille de mesures adaptées est celle des distances d'édition. Une distance d'édition mesure le coût minimal associé à une séquence d'opérations élémentaires permettant de passer d'une chaîne A à une chaîne B. L'ensemble Op des opérations élémentaires est constitué de :
- $Ajout(A, x)$, l'ajout d'un caractère x dans A.
- $Supp(A)$, la suppression d'un caractère de A.
- $Subst(A, x, y)$, la substitution dans A du caractère x par le caractère y.

Un modèle de coût, noté $coût : Op \rightarrow \mathbb{R}$, associe une valeur réelle à chacune des opérations $o \in Op$.

Soit $T_{A \rightarrow B}$, l'ensemble des séquences d'opérations $s_i = (op_1; ...; op_n)$ (avec $op_x \in Op$) permettant de passer d'une chaîne A à une chaîne B. La distance d'édition δ entre les chaînes de caractères A et B, est définie par :

$$\delta(A, B) = min_{s_i \in TA \rightarrow B} \sum_{op_x \in s_i} coût(op_x)$$

La distance d'édition associant le même coût à chacune des opérations est la distance de Levenshtein [Lev66]. D'autres modèles de coûts et algorithmes de recherche des coûts minimaux ont été proposés, notamment en bioinformatique pour des problématiques d'alignement de séquences composant les protéines ou nucléotides. On peut citer, par exemple, l'algorithme de Needleman-Wunsch [NW70] utilisant une opération de substitution associée à une matrice de similarité de caractères et une opération de gap (regroupant les opérations d'ajout et de supression) associée à un coût unitaire fixe. Une variante de cet algorithme a été proposée par [SW81].

La mesure de Jaro [Jar89] est également utilisée pour comparer deux séquences de caractères. Cette mesure est basée sur le nombre et l'ordre des caractères communs à deux chaînes. Étant donné $A = a_1...a_n$ et $B = b_1...b_m$ deux chaînes de caractères, un caractère a_i de A sera **commun** à un caractère b_j de B si $a_i = b_j$ et si $i - H \leq j \leq i + H$ (avec $H = \frac{max(card(A), card(B))}{2}$). Soit A' et B' les chaînes contenant respectivement les caractères de A communs à B et les caractères de B communs à A. Une **transposition** est définie comme les positions i telles que $a'_i \neq b'_i$. $t_{A',B'}$ représente la moitié du nombre de transpositions entre les chaînes A et B. La similarité de Jaro entre deux chaînes de caractères A et B est définie par :

$$Jaro(A, B) = 1/3. \left(\frac{card(A')}{card(A)} + \frac{card(A')}{card(B)} + \frac{card(A') - t_{A',B'}}{card(A')} \right)$$

Remarque. Les chaînes A' et B' contiennent les mêmes caractères mais dans des positions qui peuvent être différentes. Comme $card(A') = card(B')$, ces cardinalités peuvent être interchangées dans l'équation 3.4.1.

Une variante de la similarité de Jaro est celle de Jaro-Winkler [Win99] qui utilise également la taille du plus grand préfixe commun aux chaînes A et B, noté P.

$$JaroWinkler(A, B) = Jaro(A, B) + \frac{max(4, P)}{10}.(1 - Jaro(A, B))$$

En modélisant une chaîne de caractères comme une séquence de mots, les mesures ensemblistes peuvent être encore utilisées mais en considérant les mots à la place des caractères. Ces mesures sont souvent appelées distances ou similarités n-gram. Un autre moyen de comparaison consiste à utiliser le modèle vectoriel [SWY75]. Dans ce cas, les deux séquences de mots sont représentées par des vecteurs. Les dimensions des vecteurs représentent les mots constituant les deux chaînes (c.-à-d. les mots appartenant à $A \cap B$), et les composantes des vecteurs sont des poids associés aux occurrences respectives des mots dans les chaînes. Le plus souvent, le poids est donné par la mesure de TF.IDF (voir section 3.5.1). La similarité entre deux chaînes sera alors définie comme le cosinus entre les vecteurs représentant les chaînes.

Finalement, des méthodes hybrides proposent de combiner une mesure de type distance d'édition pour comparer mot à mot les deux chaînes de caractères, puis une mesure vectorielle pour agréger les similarités. SoftTF.IDF [CRF03] est un exemple de mesure hybride.

En conclusion, les mesures considérant une chaîne de caractères comme une séquence de caractères seront plus adaptées pour comparer les termes simples (identifiant, labels). L'utilisation de mesures basées sur le modèle vectoriel sera pertinente pour la comparaison de descriptions textuelles telles que les commentaires ou encore pour comparer deux entités en considérant non seulement leur identifiant mais aussi la concaténation des identifiants des entités subsumantes et/ou subsumées. Ce dernier type d'approche est considéré par [DR02] comme une approche hybride car elle fait intervenir à la fois l'information terminologique (identifiants) et l'information structurelle (subsumants et subsumés). Les méthodes syntaxiques permettent seulement de quantifier la ressemblance entre deux chaînes de caractères et sont, par conséquent, limitées à la découverte d'appariements basés sur une relation d'équivalence.

Les méthodes syntaxiques prennent en compte seulement la ressemblance physique de chaînes de caractères. Cependant, il existe dans toutes les langues des synonymes pour désigner une même entité, et l'utilisation de similarités syntaxiques sur chaînes de caractères ne permet pas de repérer de proximité dans ce cas. Afin de dépasser ces limites, d'autres méthodes basées sur des connaissances linguistiques existent.

Techniques terminologiques linguistiques

Les techniques linguistiques s'appuient sur des connaissances de la langue analysée et/ou sur des bases de données lexicales qui recensent les divers

liens sémantiques qu'entretiennent les termes (mots ou groupes de mots) d'une langue. Nous nous focalisons dans cette section sur les techniques, dites extrasèques, faisant intervenir des bases de données lexicales. Des techniques utilisées par les méthodes dites intrasèques qui exploitent les connaissances internes d'une langue, seront étudiées section 4.2.1.

Un exemple de base de données lexicale est Wordnet développée par l'université de Princeton. Les éléments de base de Wordnet sont des ensembles de termes synonymes appelés synsets. Chaque synset est associé à une définition et à un ensemble de synsets avec lesquels il est en relation. Les principales relations prises en compte pour les synsets de type nom ou groupe nominal sont :

- l'hyperonymie,
- l'hyponymie,
- la meronymie,
- l'holonymie.

En exploitant le lien de subsomption entre les termes, une base de données lexicale peut être définie comme une hiérarchie $\mathcal{L} = (C, \leq)$ où C est l'ensemble des synsets, et $\leq$ représente la relation de subsomption définie sur C. Nous notons c_0 la racine de la hiérarchie, c.-à-d. l'entité pour laquelle il n'existe pas de c_i tel que $c_0 \leq c_i$.

Les méthodes utilisant une ressource lexicale suivent, en général, le processus ci-dessous :

1. Pour chacun des deux termes t_1 et t_2 à comparer, trouver les ensembles d'entités N_{t_1} et N_{t_2} de la ressource auxquels ils correspondent.
2. Pour chaque couple $(c_i, c_j) \in N_{t_1} \times N_{t_2}$, estimer la relation qu'entretiennent c_i et c_j.
3. Déduire la relation entretenue par t_1 et t_2 à partir des relations estimées des couples (c_i, c_j).

Pour chaque couple (c_i, c_j), on peut déduire à partir de $\mathcal{L}$: $c_i \leq c_j$, $c_i \geq c_j$ ou $c_i = c_j$. Cette manière de procéder est utilisée par exemple dans [GSY04] et [RSK06]. Cependant, dans la majorité des cas, aucune relation stricte ne peut être déduite. Afin de pallier ce problème, des mesures s'appuyant sur la structure taxonomique permettent de calculer une proximité sémantique entre un couple d'entités (c_i, c_j). Les mesures proposées sont toutes dédiées à l'estimation de la relation d'équivalence et aucune mesure ne permet de quantifier l'implication ($\Rightarrow$ ou $\Leftarrow$).

Les mesures définies sur des structures hiérarchiques permettent de comparer deux entités en fonction de la quantité d'information qu'elles partagent et/ou de la quantité d'information qui leur est spécifique. Les travaux présentés dans [BKHB06] et [BHK07] ont permis de mettre en avant plusieurs moyens de modéliser la quantité d'information portée par une entité en exploitant la structure taxonomique et éventuellement les informations extensionnelles. En théorie de l'information [Sha48], la quantité d'information apportée par un évènement de probabilité p est $I(p) = -log(p)$. A partir de cela, plusieurs modèles visent à estimer la probabilité $P(c_i)$ d'une entité c_i au sein de la hiérarchie $\mathcal{L}$.

Estimateurs. Le premier estimateur disponible, proposé par [Res95], utilise un corpus de textes S. A partir de l'analyse du corpus, chaque entité c_i se

voit affectée d'une quantité n_{c_i} qui représente le nombre d'occurrences, dans le corpus S, des termes associés à c_i et à tous ses subsumés. L'estimation de $P(c_i)$ est alors définie par :

$$\widehat{P}_1(c_i) = \frac{nc_i}{nc_0}$$

Cet estimateur est utilisable dans le cas d'une hiérarchie contenant de l'héritage multiple.

Un deuxième estimateur est obtenu, sans corpus, en faisant l'hypothèse d'une distribution uniforme du nombre d'instances pour les entités d'un même niveau dans la hiérarchie. Cet estimateur est défini par :

$$\widehat{P}_2(c_i) = \frac{\widehat{P}_2(pere(c_i))}{k} = \frac{1}{k^{len(c_0,c_i)}}$$

où $pere(c_i)$ dénote le subsumant le plus spécifique de c_i, $len(c_0, c_i) = |\{c_x | c_i \leq c_x \leq c_0\}| + 1$ (longueur du chemin entre la racine c_0 et c_i) et $k > 1$ est un entier représentant le degré moyen de la hiérarchie.

Le troisième estimateur recensé dans [BHK07], consiste à faire l'hypothèse, pour une entité donnée, d'une distribution uniforme de ses instances pour ses fils. Cet estimateur est défini par :

$$\widehat{P}_3(c_i) = \frac{\widehat{P}_3(pere(c_i))}{|fils(pere(c_i))|}$$

où $fils(c_i) = \{c_x | c_x \prec c_i\}$.

Quantité d'information partagée. A partir de ces estimateurs $\widehat{P}_n(c_i)$, la quantité d'information portée par une entité c_i sera définie par :

$$I_n(c_i) = I(P_n(c_i)) = -log(\widehat{P}_n(c_i))$$

La quantité d'information commune à deux entités c_i et c_j est l'information portée par leur généralisant commun le plus spécifique. Cette quantité d'information est définie par $I_n(c_i \cap c_j) = I_n(ppg(c_i, c_j))$ où $ppg(c_i, c_j)$ est le plus petit majorant de c_i et c_j dans $\mathcal{L}$. La quantité d'information apportée conjointement par c_i et c_j est $I_n(c_i \cup c_j) = I_n(c_i) + I_n(c_j) - I_n(ppg(c_i, c_j))$. Nous pouvons remarquer que dans le cas du deuxième estimateur, $\widehat{P_2}$, la quantité d'information portée par c_i représente la longueur du chemin entre la racine c_0 et c_i ($I_2(c_i) = -log_k(\frac{1}{k^{len(c_0,c_i)}}) = len(c_0, c_i)$).

Mesures sémantiques. En combinant des schémas classiques de mesures de similarité ou de distance et les quantités d'information basées sur les estimateurs présentés ci-dessus, de nombreuses mesures de similarité entre entités issues d'une même hiérarchie sont réalisables. La table 3.1 montre pour quelques schémas de mesure (distance de Hamming, similarités de Jaccard et de Dice) et pour les trois estimateurs $\widehat{P}_1$, $\widehat{P}_2$ et $\widehat{P}_3$, les mesures correspondantes.

$\widehat{P}_1$ prend en compte un corpus d'instances associé à la hiérarchie et permet ainsi d'adapter une similarité à un contexte donné par un ensemble d'instances.

	I_1	I_2	I_3
$Q(A \cap B)$	Resnik [Res95]		
Distance de Hamming		Rada [RMBB89]	
Jaccard		Stojanovic [SMS+01] concept match (basée sur Upwards Cotopy) [MZ02]	
Dice	Lin [Lin98]	Wu et Palmer [WP94]	PSP [BKHB06]

TAB. 3.1 – Similarités sémantiques

Cependant, dans la plupart des cas un tel corpus n'est pas disponible, ainsi il sera plus facile d'utiliser $\widehat{P}_2$ ou $\widehat{P}_3$. Le dernier estimateur est celui parmi les trois proposés qui prend le plus d'information liée à la structure hiérarchique.

Concernant l'usage de l'approche basée sur une ressource terminologique, la seule méthode permettant la découverte de relation autre que l'équivalence ne s'appuie que sur une reconnaissance symbolique à partir de la ressource linguistique ([GSY04]). On peut remarquer que l'utilisation de mesures asymétriques telles que la probabilité conditionnelle (ou la confiance [AIS93]) combinées avec un des estimateurs présentés dans cette section aide à quantifier des tendances de subsomption entre entités. En effet, une mesure basée sur la probabilité conditionnelle entre deux entités c_i et c_j est définie par :

$$SUBS_{confiance}(c_i \leq c_j) = \frac{\widehat{P}_x(ppg(c_i, c_j))}{\widehat{P}_x(c_j)}$$

Cette mesure permet d'estimer la quantité d'information commune à c_i et c_j par rapport à la quantité d'information portée par c_j. Lorsque la relation $c_i \leq c_j$ est effectivement observée sur la hiérarchie alors $SUBS_{confiance}(c_i \leq c_j) = 1$. De nombreux schémas de mesures proposées dans le cadre de l'évaluation de règles d'association peuvent être utilisés (voir la section 1.2 ou les références [GH07], [Bla05], [TKS04], [HH01]).

3.4.2 Techniques structurelles

Une méthode d'alignement utilisant des critères structurels compare deux entités en s'appuyant sur la relation d'ordre dans la cadre général d'une hiérarchie. Dans le cas d'une description intensionnelle plus précise (par exemple, dans les cas de schémas de bases de données, XML, ou objets et d'ontologies OWL), elle pourra également utiliser des critères concernant les attributs ou/et les relations transversales entre entités.

[SE05] distingue deux types de méthodes structurelles : celles basées sur la structure interne d'une entité et celles basées sur la structure externe. La structure interne d'une entité représente les attributs possédés par l'entité (attributs

de type simple, cardinalités, restrictions). La structure externe représente les relations qu'entretient l'entité avec d'autres. Au niveau de la structure externe, on distinguera les méthodes basées sur la relation d'ordre, des méthodes basées sur tous les autres types de relations.

Les méthodes structurelles externes utilisent généralement un alignement préalable A (filtré ou non) et permettent de le raffiner. Ce type de méthode est ainsi utilisé dans un schéma de composition linéaire (section 3.3.2). L'idée sous-jacente à une méthode structurelle est la suivante : deux entités issues de hiérarchies distinctes seront en relation si les entités de leur voisinage ont tendance à être en relation. Le voisinage d'une entité est défini comme l'ensemble des entités reliées par un chemin de longueur déterminée dans le graphe de la relation considérée (relation d'ordre ou autres relations transversales).

Nous nous intéresserons dans cette section surtout aux approches structurelles basées sur la relation d'ordre. Nous présenterons tout de même brièvement les principes des approches basées sur les attributs. Nous donnerons également un aperçu des méthodes structurelles se basant sur n'importe quel type de relations.

Basées sur les attributs (structurelles internes)

La comparaison basée sur la structure interne des entités est utilisée dans les cas où les entités ont des définitions intentionnelles précises. Ces méthodes sont couramment utilisées pour aligner des schémas de bases données ou encore des ontologies. Cependant, elle ne sont que peu adaptées à des structures ayant des définitions simples telles que les catalogues Web, les répertoires Web, ou encore les thésaurus.

Le principe sous-jacent à ce type de méthode est d'utiliser une table de similarités afin de quantifier le degré de compatibilité entre deux types d'attributs ou pour comparer les cardinalités, restrictions et autres contraintes d'intégrité. Cette approche a été tout d'abord utilisée par des méthodes d'alignement et d'intégration de schémas de bases de données [MBR01], [CAdV01], [PSU98, PTU03] puis reprise dans le contexte d'alignement d'ontologies [CFM05], [TLL+06].

Les méthodes basées sur les attributs utilisent ces similarités soit en les agrégeant (par une des techniques présentées section 3.3.1) pour déterminer une similarité entre un couple d'entités, soit en entrée d'un algorithme structurel de propagation de similarités (Cupid [MBR01], Similarity Flooding [MGMR02]).

Basées sur la relation d'ordre

En s'appuyant sur la relation d'ordre, une entité x d'une hiérarchie peut avoir trois types de voisinage :

- ses subsumants, définis par $N_>(x) = \{x' | x < x'\}$,
- ses subsumés, définis par $N_<(x) = \{x' | x' < x\}$,
- ses soeurs, définies par $N_s(x) = \{x' | x \prec x'' \wedge x' \prec x'' \wedge x' \neq x\}$

Les méthodes structurelles basées sur la relation d'ordre utilisent ces voisinages (soit en totalité, soit en partie) et un alignement $A = (V, q)$ fourni en

entrée. A partir de ces informations, elles ont pour but de quantifier la possibilité d'appariement de deux entités x et y en s'appuyant sur les éléments de leurs voisinages respectifs qui sont eux-mêmes contenus dans l'alignement d'entrée. Deux alternatives, utilisant différents niveaux d'information sont possibles : la première consiste à utiliser seulement le nombre d'entités de leurs voisinages qui sont en correspondance dans V ; la deuxième alternative considère en plus les valeurs de qualités associées aux éléments de correspondance concernés.

Etant donné une relation $\mathcal{R}$ et un alignement $A = (V, q)$, $N_>(x) \cap N_>(y)$ représente l'ensemble des couples de $N_>(x) \times N_>(y)$ qui sont dans V, c.-à-d. l'ensemble des éléments $x'\mathcal{R}y' \in V$ où $x' \not> x$ et $y' \not< y$.

$$N_>(x) \cap N_>(y) = \{(x', y', \mathcal{R}) | x' \in N_>(x) \wedge y' \in N_>(y) \wedge (x, y, \mathcal{R}) \in V\}$$

De manière analogue, nous définissons les ensembles $N_<(x) \cap N_<(y)$ et $N_s(x) \cap N_s(y)$.

Si la cardinalité de l'alignement A est $0,1 - 0,1$ alors le nombre maximal d'éléments de $N_>(x) \cap N_>(y)$ sera $max(|N_>(x)|, |N_>(y)|)$. Dans ce cas, à partir des cardinalités des ensembles $N_>(x)$, $N_>(y)$ et $N_>(x) \cap N_>(y)$, n'importe quelle mesure de similarité ensembliste (présentées section 3.1.2) peut être utilisée [4] :

$$s(x, y) = \frac{|N_>(x) \cap N_>(y)|}{m_\alpha(|N_>(x)|, |N_>(y)|)}$$

Cependant, si l'alignement d'entrée, A, est de cardinalité $0,n - 0,n$, alors ces mesures de similarités ne pourront pas être appliquées étant donné qu'elles normalisent le résultat par une moyenne des cardinalités des ensembles $N_>(x)$ et $N_>(y)$. Dans ce cas, on utilise la cardinalité du produit cartésien $N_>(x) \times N_>(y)$ pour normaliser. La similarité sera alors définie ainsi :

$$s(x, y) = \frac{|N_>(x) \cap N_>(y)|}{|N_>(x) \times N_>(y)|}$$

La sémantique des similarités calculées par ce type d'approche sera la suivante : une entité x sera en relation $\mathcal{R}$ avec l'entité y si une relativement grande quantité de leurs entités subsumantes (respectivement subsumées ou soeurs) sont en relation $\mathcal{R}$ dans l'alignement d'entrée A.

La deuxième alternative consiste à faire également intervenir les valeurs de qualité associées par q (si les valeurs de q sont comprises dans l'intervalle $[0-1]$) dans l'alignement A. Dans ce cas, la quantité $|N_>(x) \cap N_>(y)|$ sera estimée par $P(|N_>(x) \cap N_>(y)|)$:

$$P(|N_>(x) \cap N_>(y)|) = \sum_{(x', y', \mathcal{R}) \in N_>(x) \cap N_>(y)} q(x', y', \mathcal{R})$$

Dans le cas des ensembles d'entités subsumantes ou subsumées, on peut également restreindre les ensembles uniquement aux descendants (subsumés) ou parents (subsumants) directs.

[4] Dans le cas de la mesure de Jaccard, la quantité $|N_>(x) \cup N_>(y)|$ sera égale à $|N_>(x)| + |N_>(y)| - |N_>(x) \cap N_>(y)|$

On remarquera que les mesures s'appuyant sur les subsumants communs, $N_>(x) \cap N_>(y)$, donnent les mêmes valeurs pour les couples de soeurs de x et y. Cette remarque est également valable, pour les mesures basées sur l'ensemble des soeurs communes $N_s(x) \cap N_s(y)$, lorsque $(x, y, \mathcal{R}) \notin V$, pour tous les couples de soeurs qui ne sont également pas présents dans V. C'est pourquoi ces mesures doivent être en principe combinées avec la mesure prenant en compte les descendants communs à x et y.

Basées sur les relations transversales

Sur des schémas de bases de données où la relation d'ordre n'est pas forcément définie, des méthodes permettant de prendre en compte des relations entre entités de manière générale ont été proposées. Ces méthodes sont également utiles pour des schémas possédant de nombreuses relations (schémas objets, XML, ontologies). Dans ces cas, un schéma est constitué d'un ensemble d'entités C et d'un ensemble de relations binaires $\mathcal{P}_c$ définies sur C. Chaque relation $p_i \in \mathcal{P}_c$ peut être représentée par un graphe orienté $G_{p_i} = (C, p_i)$ où C sera l'ensemble des sommets et p_i l'ensemble des arcs.

Parmi les méthodes proposées, Similarity Flooding (SF) [MGMR02] permet de prendre en compte chaque relation p_i séparément. Cet algorithme permet de raffiner les valeurs de qualité données par un alignement $A(V, q)$ de manière itérative. A chaque itération, SF recalcule la valeur de qualité d'une correspondance $(x, y, \Leftrightarrow)$ de V en considérant les correspondances $(x'_j, y'_j, \Leftrightarrow)$ issues du voisinage immédiat de x et de y dans les graphes de la relation p_i (et de sa réciproque p_i^{-1}).

La méthode Cupid [MBR01] est également munie d'un algorithme d'alignement structurel, appelé treeMatch, permettant d'aligner des structures hiérarchiques. Cependant le calcul de la similarité structurelle est basé sur des compatibilités entre les types de données des attributs (similarité structurelle interne). Cette méthode est ainsi restreinte à des schémas de bases de données, XML ou objets.

Une autre méthode, ANCHOR-PROMPT [NM01], [NM03], a été proposée dans le but d'aligner des ontologies. A partir d'entités syntaxiquement proches données dans un alignement $A(V, q)$ (filtré) et des graphes globaux $G_{\mathcal{P}_c}$ associés à chaque ontologie, cette méthode va considérer toutes les paires de chemins reliant les entités de V. L'algorithme détermine la similarité entre deux entités à partir de leur fréquence d'apparition dans des positions identiques dans les paires de chemins.

La méthode GMO [HaYQW05] est également une méthode structurelle considérant tout type de relations (relation d'ordre et propriétés). Cette méthode représente des ontologies RDFS/OWL en utilisant la sémantique de triplet RDF (Sujet, Prédicat, Objet). L'ensemble des triplets est représenté sous forme d'un graphe biparti [HG04]. Les deux sous-ensembles de sommets de ce graphe biparti sont, d'une part, les sujets, les objets et les prédicats et d'autre part, les sommets représentant les triplets. Dans le premier ensemble de sommets (celui contenant les sujets, objets et prédicats), deux sortes de sommets sont distinguées :

- Les entités internes, définies dans l'ontologie,
- Les entités externes, représentant les éléments du langage ou les entités issues d'une autre ontologie.

Pour mesurer la similarité entre les entités issues de deux ontologies représentées par leur graphe biparti, GMO utilise un algorithme itératif de similarité structurelle. Le principe est le suivant : deux triplets (issus respectivement des deux ontologies) seront similaires si les entités externes auxquelles ils sont reliés ont les mêmes rôles. A partir de cette première similarité, deux entités internes seront similaires si les triplets auxquelles elles sont reliées (avec le même rôle) sont similaires.

3.5 Les techniques d'alignement extensionnelles

La description extensionnelle d'une hiérarchie $\mathcal{H} = (C, \leq, \mathcal{A}, O, \sigma)$ est composée des objets O associés aux entités C par σ. L'utilisation de l'information extensionnelle (des instances) est intéressante lorsque les informations intensionnelles sont limitées [RB01]. De manière générale, les données semi-structurées, les répertoires Web, les catalogues de boutiques en ligne, les thésaurus associés à des documents, les forums de discussions sont des exemples typiques de structures ayant beaucoup d'information extensionnelle mais possédant un schéma limité aux noms des entités et à leur structuration hiérarchique.

Une méthode extensionnelle utilise pour l'alignement de deux hiérarchies $\mathcal{H}_1 = (C_1, \leq, \mathcal{A}_1, O_1, \sigma_1)$ et $\mathcal{H}_2 = (C_2, \leq, \mathcal{A}_2, O_2, \sigma_2)$ les objets associés à chaque entité par la relation d'indexation σ_i, $i \in \{1, 2\}$. Le principe de ces méthodes est d'induire la relation éventuelle qu'entretiennent deux entités $x \in C_1$ et $y \in C_2$ en s'appuyant sur leurs extensions respectives $\sigma_1(x)$ et $\sigma_2(y)$.

Le moyen intuitif de détecter une éventuelle relation entre deux entités x et y consiste à comparer leurs extensions. Il existe 4 cas possibles :

- $\sigma_1(x) = \sigma_2(y)$ alors $x \Leftrightarrow y$
- $\sigma_1(x) \subseteq \sigma_2(y)$ alors $x \Rightarrow y$
- $\sigma_2(y) \subseteq \sigma_1(x)$ alors $x \Leftarrow y$
- $\sigma_1(x) \not\subseteq \sigma_2(y)$ et $\sigma_2(y) \not\subseteq \sigma_1(x)$ alors x n'est pas en relation avec y

Cependant, une première contrainte rend ce principe simple non applicable tel quel. En effet, il est rare que deux hiérarchies partagent les mêmes extensions ($O_1 \neq O_2$). Afin de résoudre le premier problème, les approches extensionnelles réalisent un pré-traitement sur les hiérarchies afin de les rendre comparables. Il existe trois approches possibles :

- Réduire leur extension (et leur relation d'indexation) à $O_1 \cap O_2$.
- Augmenter leur extension (et leur relation d'indexation) à $O_1 \cup O_2$.
- Extraire une autre représentation des extensions et donc une autre relation d'indexation.

La première approche consiste à réduire les données extensionnelles. Pour chaque hiérarchie $\mathcal{H}_i$, son extension O_i ($i \in \{1, 2\}$) devient $O_1 \cap O_2$. Pour toute entité $c \in C_i$, la relation d'indexation σ_i est remplacée par la relation σ'_i :

$$\sigma'_i(c) = \sigma_i(c) \cap O_j, j \in \{1, 2\} - \{i\}$$

Cette approche (illustrée figure 3.4) est la plus simple à mettre en place, mais

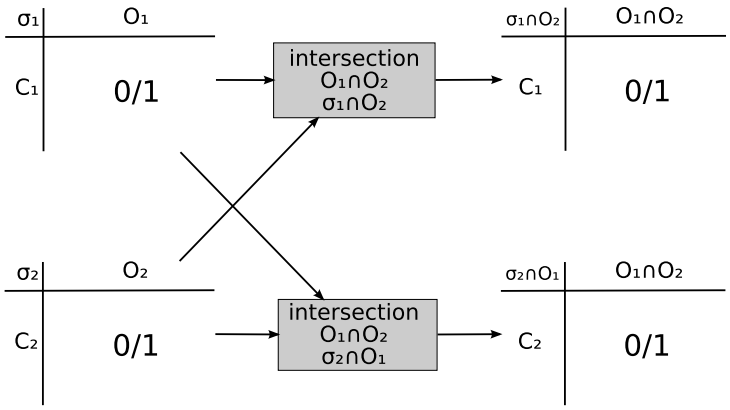


FIG. 3.4 – Intersection

il est nécessaire que l'intersection $O_1 \cap O_2$ ne soit pas égale à l'ensemble vide et souhaitable qu'elle soit relativement conséquente afin que les résultats obtenus soit statistiquement valides. De ce fait, ce type d'approche est rarement utilisé tel quel.

La deuxième approche consiste à étendre l'extension de chaque hiérarchie sur l'union des objets $O_1 \cup O_2$ et, par conséquent, étendre chaque relation d'indexation aux objets de l'autre hiérarchie. Afin de réaliser l'association de nouveaux objets aux entités d'une hiérarchie, les méthodes s'appuyant sur ce type d'approche doivent construire un modèle d'indexation. La construction du modèle d'indexation s'appuie sur des techniques de classification supervisée. La classification supervisée est une approche d'apprentissage automatique permettant de construire une fonction associant une classe à un objet donné [Mit97]. La fonction de classification est induite par l'apprentissage d'un modèle à partir d'exemples d'objets associés à la classe désirée.

La dernière approche consiste à changer complètement l'extension et par conséquent la relation d'indexation. Ce changement de représentation est fait par l'extraction et la sélection de descripteurs issus des données extensionnelles. Ces descripteurs peuvent être des termes ou mots dans le cas de données textuelles (catalogues et répertoires Web, forums, thésaurus) ou encore des valeurs d'attributs dans le cas de données structurées (ontologies, schémas de bases de données ou objets).

Une fois cette première étape de prétraitement réalisée, la comparaison devient alors possible. Cependant, il existe un second problème concernant la comparaison stricte des extensions. En effet, il est courant de constater quelques contre-exemples à l'égalité ou l'inclusion stricte sans que la tendance générale ne puisse être contestée. Ainsi, les méthodes ont généralement recours à des mesures permettant de quantifier la qualité de la tendance observée. Nous pouvons remarquer que la grande majorité des méthodes extensionnelles sont basées sur l'utilisation de similarités et que l'utilisation de mesures asymétriques parmettant de qualifier les tendances implicatives (voir section 1.2.2) n'a pas été étudiée dans ce contexte.

Nous détaillons dans cette section les méthodes de classification des instances qui permettent de classer les instances de chaque hiérarchie dans l'autre. Ensuite, nous présentons les méthodes de caractérisation des instances qui permettent de redéfinir l'extension de chaque entité par un ensemble de descripteurs issus de leurs instances (mots-clés, statistiques, ...). Finalement, nous discuterons des mesures pouvant être utilisées dans le but de comparer et d'induire la relation entretenue par deux entités.

3.5.1 Augmentation de l'extension par classification supervisée

Les méthodes s'appuyant sur la classification vont permettre d'augmenter les relations d'indexation σ_i ($i \in \{1, 2\}$) de deux hiérarchies $\mathcal{H}_1$ et $\mathcal{H}_2$ sur la réunion $O_1 \cup O_2$ de leurs ensembles d'instances. A l'issue de cette étape de prétraitement, chaque hiérarchie $\mathcal{H}_i$ sera redéfinie par $\mathcal{H}'_i = (C_i, \leq, O_1 \cup O_2, \sigma'_i)$. Les relations d'indexation augmentées σ'_i sont définies de la manière suivante :

$$\sigma'_i(c) = \sigma_i(c) \cup \mu_i(c)$$

où $\mu_i(c)$ représente les objets de O_j associés à l'entité $c \in C_i$. Chaque relation μ_i est construite grâce à un algorithme de classification supervisé.

De part la structuration hiérarchique des entités, le processus de classification devra associer chaque objet de O_j à plusieurs entités de C_i. De plus, les objets de O_i, servant d'exemples d'apprentissage, sont également associés à plusieurs entités de C_i par la relation σ_i. Ce problème de classification est donc multi-classes. Afin de pouvoir utiliser des algorithmes d'apprentissage classiques, on peut utiliser la technique classique du un-contre-tous. Ainsi, chaque entité c aura sa propre fonction de classification binaire H_c. Cette fonction utilisera $\sigma_i(c)$ comme exemples positifs et $O_i - \sigma_i(c)$ comme exemples négatifs. La classe prédite pour un objet $o \in O_j$ pourra être c ou bien $\overline{c}$.

La relation d'indexation μ_i sera donc définie à partir des fonctions de classification H_c (où $c \in C_i$) par :

$$\mu_i(c) = \{o \in O_j | H_c(o) = c\}$$

Pour résumer, le processus d'augmentation des relations d'indexation (illustré figure 3.5), réalisé sur chaque hiérarchie $\mathcal{H}_i$ ($i \in 1, 2$), se déroule en 3 étapes successives :

1. Apprentissage à partir des fonctions H_c à partir des relations σ_i et des ensembles O_i.
2. Construction des ensembles μ_i à partir des fonctions H_c.
3. Union des ensembles d'instances en $O_1 \cup O_2$ et des relations σ_i et μ_i en σ'_i.

Les méthodes s'appuyant sur une étape de classification ont souvent été conçues pour fonctionner sur des objets de type documents textuels, elles utilisent généralement un modèle de prédiction développé dans le contexte de la classification automatique de documents. Nous présentons dans cette section

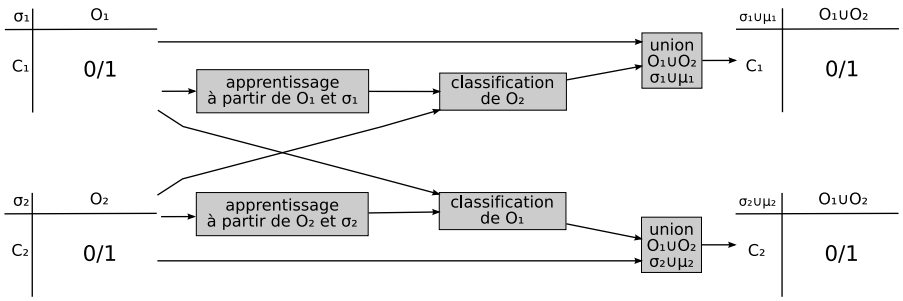


FIG. 3.5 – Union - Classification

deux méthodes : une probabiliste, s'appuyant sur le modèle bayésien-naïf et une méthode vectorielle, le prédicteur de Rocchio ou TF/IDF.

Ces méthodes représentent chaque objet $d \in O_i$ par un ensemble de descripteurs $\{t1, ..., t_{|d|}\}$. A partir de cette représentation, elles permettent de calculer une fonction de confiance $F(c|d)$ d'association de l'objet d à une entité c. A partir de cette fonction de confiance, la fonction de prédiction peut suivre la règle de décision du Maximum A Posteriori (MAP) afin de déterminer la classe c ou $\overline{c}$ à associer à un objet d :

$$H(d) = argmax_{x \in \mathcal{C}_c} F(x|d) \tag{3.5}$$

$\mathcal{C}_c$ représente l'ensemble de classes possibles pour d. Dans notre cas, pour chaque entité c, les classes possibles seront $\mathcal{C}_c = \{c, \overline{c}\}$.

Une méthode probabiliste : le prédicteur bayésien-naïf

Le principe d'un prédicteur bayésien-naïf est d'estimer la probabilité $P(x|d)$ qu'un objet $d \in O_j$ soit associé à une classe $x \in \mathcal{C}$.

$$F_{BAYES}(x|d) = Pr(x|d) \tag{3.6}$$

En appliquant le théorème de Bayes à $Pr(x|d)$, on obtient :

$$Pr(x|d) = \frac{Pr(x)Pr(d|x)}{\sum_{x' \in \mathcal{C}} Pr(d|x')} \tag{3.7}$$

Une estimation $\widehat{Pr}(x)$ de $Pr(x)$ peut être facilement calculée à partir de l'ensemble d'apprentissage.

$$\widehat{Pr}(x) = \frac{|x|}{|O_i|}$$

Le modèle utilisé fait l'hypothèse simplificatrice qu'une occurrence d'un mot est uniquement dépendante de l'entité à laquelle il est associé, c.à-d. que cette

occurrence est indépendante des autres mots dans le document. A partir de cette hypothèse, la probabilité $Pr(d|c)$ est définie ainsi :

$$Pr(x|c) \approx \prod_{k=1}^{|d|} Pr(t_k|x) \tag{3.8}$$

L'estimation de $Pr(d|x)$ est réduite à estimer chaque $Pr(t_k|x)$ de manière indépendante. Une estimation de $Pr(t_k|x)$ peut être donnée par l'estimateur de Laplace :

$$\widehat{Pr}(t_k|x) = \frac{1 + TF(t_k, x)}{|F| + \sum_{t' \in F} TF(t', x)}$$

En combinant les équations 3.6, 3.7 et 3.8 , la probabilité d'associer un document d à l'entité c est finalement :

$$Pr(x|d) = \frac{Pr(x).\prod_{k=1}^{|d|} Pr(t_k|x)}{\sum_{x' \in \mathcal{C}} Pr(x').\prod_{k=1}^{card(d)} Pr(t_k|x')} \tag{3.9}$$

Etant donné que le dénominateur est indépendant de l'entité considérée, la fonction de classification H_{BAYES} d'un objet d est définie par :

$$H_{BAYES}(d) = argmax_{x \in \mathcal{C}} Pr(x).\prod_{k=1}^{|d|} Pr(t_k|x)$$

Une version améliorée du prédicteur de Bayes a été proposée afin de prendre en compte dans le processus d'apprentissage, l'information issue de chacune des hiérarchies [AS01]. L'intuition sous-jacente à leur proposition est que si deux objets sont associés à la même entité dans $\mathcal{H}_1$ alors il est préférable qu'ils soient associés à la même entité de $\mathcal{H}_2$ (et vice-versa). Cette intuition se traduit par la prise en compte de l'entité source $s \in C_i$ à laquelle un document $d \in O_i$ est associé. Ainsi la probabilité $Pr(x|d, s)$ qu'un document d soit associé à une entité x sachant son indexation d'origine s s'écrira :

$$Pr(x|d, s) = \frac{Pr(x)Pr(d, s|x)}{\sum_{x' \in \mathcal{C}} Pr(d, s|x')}$$

En supposant l'indépendance de d et s étant donné x, la probabilité peut s'écrire :

$$Pr(x|d, s) = \frac{Pr(x)Pr(d|x)Pr(s|x)}{\sum_{x' \in \mathcal{C}} Pr(d, s|x')}$$

Et puisque $Pr(x|s)Pr(s) = Pr(s|x)Pr(x)$, on obtient finalement :

$$Pr(x|d, s) = \frac{Pr(x|s)Pr(d|c)}{\sum_{x' \in \mathcal{C}} Pr(d|s, x')} \tag{3.10}$$

Le seul changement significatif entre les équations 3.9 et 3.10 est la probabilité $Pr(x)$ qui est remplacée par $Pr(x|s)$. La somme de probabilités apparaissant au dénominateur change elle aussi mais comme elle est identique pour toute les

entités, son résultat n'intervient pas dans la décision. Une première estimation possible de $Pr(x|s)$ peut être faite en utilisant les résultats donnés par un prédicteur bayésien-naïf :

$$\widehat{Pr}(x|s) = \frac{\textit{Nombre de documents de } s \textit{ classés dans } x}{|\sigma(s)|}$$

Cependant, comme cette prédiction dépend entièrement des résultats produits par le prédicteur bayésien-naïf, [AS01] ont proposé de combiner les informations initiales portées par $Pr(x)$ et les informations du prédicteur bayésien-naïf pondérées par un coefficient $\omega \leq 0$. Lorsque $\omega = 0$, alors le prédicteur amélioré sera identique au bayésien-naïf. La nouvelle estimation de $Pr(x|s)$ devient alors :

$$\widehat{Pr}(x|s) = \frac{|x| \times (\textit{Nombre de documents de } s \textit{ classés dans } x)^{\omega}}{|O_i| \times |\sigma(s)|^{\omega}}$$

Le résulats présentés dans [AS01] montrent une amélioration relative de la classification de 25% à 30%.

Une méthode vectorielle : le prédicteur de Rocchio

Le modèle vectoriel introduit par [SWY75] est un modèle d'algèbre linéaire couramment utilisé en recherche et indexation d'information. Il permet de représenter des documents textuels par des vecteurs dans un espace multi-dimensionnel où les dimensions désignent des termes (mots-clés). Cette méthode vectorielle représente donc chaque objet d par un vecteur $\overrightarrow{d} = (d_{t_1}, ...d_{t_{|T|}})$ où T est l'ensemble des descripteurs issus des objets (dans le cas de documents ce sont des mots). Une composante d_{t_i} représente le poids du descripteur t_i dans le document d. Ce poids est donné par la mesure $TF.IDF$ définie ainsi :

$$TF.IDF(t,d) = TF(t,d).IDF(t)$$

$TF(t,d)$ (Term Frequency) représente la fréquence d'un terme t dans un document d.

$IDF(t)$ (Inverse Document Frequency) est fonction de l'inverse du nombre de documents dans lesquels le terme t apparaît qui donné par $DF(t)$:

$$IDF(t) = log\left(\frac{|O|}{DF(t)}\right)$$

L'idée sous-jacente à la mesure $TF.IDF$ est qu'un descripteur t sera important pour un document d s'il apparaît fréquemment dans ce document (TF élevé) et si ce descripteur n'apparaît pas dans beaucoup de documents (DF faible).

Ensuite, pour chaque entité c, on construit un vecteur caractéristique $\overrightarrow{c} = (c_{t_1}, ...c_{t_{|F|}})$ où chaque coordonnée c_{t_j} représente la différence pondérée entre la moyenne des coordonnées normalisées des objets associés à l'entité (exemples) et celle des objets non associés à l'entité (contre-exemples).

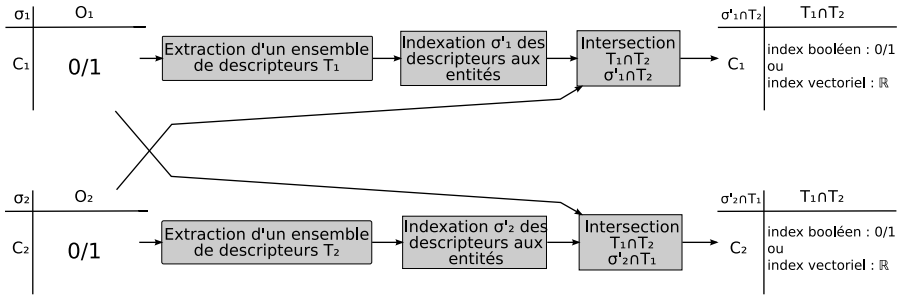


FIG. 3.6 – Processus de redéfinition d'indexation

$$c_{t_j} = \alpha \frac{1}{|c_j|} \sum_{d \in c_j} \frac{d_{t_j}}{\| \overrightarrow{d} \|} - \beta \frac{1}{|C - c_j|} \sum_{d \in C - c_j} \frac{d_{t_j}}{\| \overrightarrow{d} \|}$$

A partir du vecteur d'un objet d et des vecteurs caractéristiques de chaque entité c, le prédicteur de Rocchio utilise la mesure de cosinus entre ces deux vecteurs. Un document d sera donc associé à l'entité pour laquelle le cosinus entre leurs deux vecteurs respectifs est maximisé :

$$H_{Rocchio}(d) = argmax_{c_j \in C} cos(\overrightarrow{c_j}, \overrightarrow{d})$$

où

$$cos(\overrightarrow{c_j}, \overrightarrow{d}) = \frac{\overrightarrow{c_j} . \overrightarrow{d}}{\| \overrightarrow{c_j} \| . \| \overrightarrow{d} \|}$$

3.5.2 Réindexation des données extensionnelles

Cette famille de méthodes permet de redéfinir complétement les extensions des entités par un ensemble de descripteurs extraits et sélectionnés à partir de l'analyse du contenu des instances. Cette analyse va permettre de mettre en évidence un certain nombre de descripteurs communs aux deux hiérarchies à aligner. Ces descripteurs vont être indexés aux entités des hiérarchies qui pourront être ainsi comparées grâce à cette nouvelle représentation.

Pour les deux hiérarchies $\mathcal{H}_i = (C_i, \leq, \mathcal{A}_i, O_i, \sigma_i)$, ce processus (illustré figure 3.6) se déroule en trois étapes successives :

1. Extraction et sélection des ensembles de descripteurs T_i pour chaque ensemble d'instances O_i.
2. Association des entités C_i aux descripteurs T_i par une nouvelle relation d'indexation γ.
3. Réduction des ensembles de descripteurs à l'intersection $T_1 \cap T_2$, et des relations d'indexations en γ'_i.

Finalement, chaque hiérarchie $\mathcal{H}_i$ sera redéfinie en hiérarchie $\mathcal{H}_i = (C_i, \leq, \mathcal{A}_i, T_1 \cap T_2, \gamma'_i)$ conservant la même structure, mais indexant un nouvel ensemble de données.

Deux types d'indexation sont utilisés : (1) l'indexation booléenne où chaque entité $c \in C$ est associée par la relation γ à un sous-ensemble de descripteurs $\gamma(c) \subseteq T$; (2) l'indexation vectorielle où chaque entité $c \in C$ est associée par γ à un vecteur $\overrightarrow{v_c} = \gamma(c)$ de dimension $|T|$ où chaque dimension représente un descripteur et les composantes sont des valeurs quantifiant le degré de représentativité du descripteur pour l'entité donnée.

L'intérêt d'une indexation booléenne réside surtout dans sa simplicité de représentation, plus intelligible à l'utilisateur qui pourra facilement comprendre les descripteurs associés à une entité. Les indexations vectorielles sont des modèles plus évolués qui donnent de meilleurs résultats en recherche d'information [GF98]. Néanmoins, la représentation d'une entité par un vecteur est plus difficile à appréhender par l'utilisateur. De plus, contrairement à l'indexation vectorielle, l'indexation booléenne permet de conserver une représentation ensembliste des descripteurs associés à une entité par la relation γ, et ainsi le morphisme entre la relation d'ordre et l'inclusion des ensembles de descripteurs. Dans le cadre de l'alignement, toutes les méthodes proposées s'appuyant sur la redéfinition de l'indexation ([LG01], [HYNT04], [QHC06]) utilisent un modèle vectoriel.

Parmi les méthodes se basant sur l'indexation vectorielle, [LG01] proposent de construire les vecteurs caractéristiques de chaque entité $c_i \in C$ en utilisant la méthode Rocchio (voir section 3.5.1, formule 3.5.1).

La méthode Semantic Category Matching Approach (SCM) [HYNT04] propose une extraction et sélection des descripteurs basées sur la divergence de Kullback-Leibler. L'association entre un descripteur t_x et une entité $c_i \in C$, sera évaluée par l'entropie relative (divergence de Kullback-Leibler) entre la probabilité que le descripteur t_x apparaisse dans les objets associés à l'entité c_i par rapport à la probabilité moyenne que le descripteur apparaisse dans les objets associés aux entités soeur de c_i.

$$KL(t_x, c_i) = DF(t_x, c_i).log\left(\frac{DF(t_x, c_i)}{\frac{1}{|soeurs(c_i)|}\sum_{c_j \in soeurs(c_i)} DF(t_x, c_j)}\right)$$

Les valeurs calculées pour chaque entité c_i sont ensuite propagées à toutes ses entités subsumantes. Finalement, chaque entité c_i sera associée à un vecteur caractéristique dont les descripteurs seront sélectionnés par rapport à leur rang (les auteurs de SCM proposent de sélectionner les 30 descripteurs les mieux valués).

3.5.3 Comparaison d'extension

Une fois le prétraitement réalisé, les deux hiérarchies ont des extensions comparables. Afin de déduire la relation éventuelle que peuvent entretenir deux entités $x \in C_1$ et $y \in C_2$, les méthodes comparent leur extensions $\sigma(x)$ et $\sigma(y)$ au moyen d'une mesure dont le résultat sera noté $m(\sigma(x), \sigma(y))$. Parmi les mesures utilisables, on peut distinguer les similarités (ou distances) et les mesures asymétriques permettant d'évaluer la qualité des règles d'association. Les premières mesures, qui sont symétriques, permettent seulement d'estimer la qualité de la relation $x \Leftrightarrow y$. Les secondes mesures permettent également

d'évaluer la relation $x \Rightarrow y$ (ou $x \Leftarrow y$).

Parmi toutes les méthodes extensionnelles proposées dans la littérature, aucune ne s'appuie réellement sur des mesures asymétriques permettant de qualifier la tendance implicative entre deux entités. On peut tout de même noter une tentative intéressante proposée par [DMDH04]. En effet, dans ce papier les auteurs mentionnent la possiblité d'utiliser une mesure asymétrique, appelée MSP (Most Specific Parent). Cette mesure est définie par :

$$MSP(x \rightarrow y) = \begin{cases} \frac{|\gamma_1(x) \cap \gamma_2(y)|}{|\gamma_2(y)|} \ si\ \gamma_1(x) \subseteq \gamma_2(y) \\ 0\ sinon \end{cases}$$

Cependant cette mesure permet seulement de trouver l'implication stricte ayant la conclusion (l'entité cible) la plus spécifique pour une prémisse (entité source) donnée. En effet, cette mesure ne permet pas de quantifier la qualité d'une quasi-implication puisqu'elle est égale à 0 si $\gamma_1(x) \not\subseteq \gamma_2(y)$.

En fonction du type de prétraitement qu'effectue une méthode d'alignement, les mesures utilisées peuvent être spécialisées. En effet, les méthodes de réindexation de l'extension, typiquement basées sur le modèle vectoriel, utilisent, naturellement, la mesure de cosinus pour évaluer la ressemblance entre deux vecteurs représentant les entités.

Les méthodes basées sur l'augmentation de l'extension par classification supervisée, sont quant à elles plus indépendantes vis-à-vis de la mesure utilisée. En effet, dans ce cas la comparaison des extensions revient à comparer deux ensembles, et par conséquent, n'importe quelle mesure ensembliste est utilisable [DMDH04].

3.6 Comparaison de méthodes d'alignement

Nous proposons une comparaison synthétique d'une vingtaines de méthodes d'alignement. Cette comparaison est divisée en trois parties :

- comparaison basée sur les caractéristiques globales ;
- comparaison des techniques intensionnelles ;
- comparaison des techniques extensionnelles.

3.6.1 Caractéristiques globales

La table 3.2 décrit les méthodes sélectionnées selon leurs entrées et sorties. Au regard des types de schémas, nous avons sélectionné majoritairement des méthodes d'alignement destinées aux hiérarchies textuelles (H.T.) et aux ontologies (RDFS/OWL). Nous considérons, tout de même, deux méthodes qui permettent d'aligner des structures de type schémas relationnel (Rel.), orientés objet (O.O.) ou XML. En effet, les méthodes Cupid [MBR01] et Similarity Flooding [MGMR02] sont, d'une part, souvent citées dans la littérature traitant de l'alignement d'ontologies, et d'autre part, des méthodes représentatives des approches structurelles génériques qui peuvent être utilisées dans ce contexte. La méthode COMA[5] [DR02] a été conçue pour obtenir un maximum de flexibilité

[5] Du fait de sa nature modulaire, cette méthode est plus une « méta-méthode » d'alignement.

Méthode	**types de schémas**	**rel.**	**cardinalité**
ANCHOR-PROMPT [NM01]	RDFS/OWL	⇔	0-n,0-n
ASCO1 [BDKG04]	RDFS	⇔	0,n-0,n
AUTOMS (H-CONE merge) [KVS06]	RDFS/OWL	⇔	?
CAIMAN [LG01]	H.T.	⇔	0,1-0,1
COMA [DR02]	tous types de schémas	⇔	0,1-0,1 0,1-0,n 0,n-0,n
Cupid [MBR01]	XML, O.O.	⇔	0,n-0,n
GLUE [DMDH02]	H.T.	⇔	
GMO [HaYQW05]	RDFS/OWL	⇔	
V-Doc [QHC06]	RDFS/OWL	⇔	
H-match [CFM05]	RDFS/OWL	⇔	0,1-0,1 0,1-0,n
OLA [EV04]	RDFS/OWL	⇔	0,1-0,1
oMap [ST05]	RDFS/OWL	⇔	0,1-0,n
OplMap [NS06]	H.T.	⇔	0,n-0,1
QOM [ES04]	RDFS/OWL	⇔	0,1-0,1
RiMOM [TLL+06]	H.T., RDF-S/OWL	⇔	0,1-0,1 0,1-0,n
SBI - Hical [IHT04]	H.T.	⇔	0,n-0,n
SCM [HYNT04]	H.T., RDF-S/OWL	⇔	0,1-0,n
SF [MGMR02]	Rel., XML, O.O., RDF	⇔	0,1-0,1
S-match/ Ctx-match [GSY04]	H.T., OWL/RDFS	⇔, ⇒,⊓	

TAB. 3.2 – Comparaison des méthodes par rapport à leurs entrées et sorties

Méthode	**combinaison**	**sélection**	**post-traitements**
ANCHOR-PROMPT [NM01]	linéaire (utilise alignement d'entrée)	seuillage	-
ASCO1 [BDKG04]	moyenne pondérée (à deux niveaux : terminologique et structurel)	seuillage	
AUTOMS (H-CONE merge) [KVS06]	-	maxi. locale	-
CAIMAN [LG01]		maxi. locale	-
COMA [DR02]	statistique au choix (max, moyenne...)	au choix (maxi. locale, seuillage)	
Cupid [MBR01]	moyenne pondérée	seuillage	-
GLUE [DMDH02]	moyenne pondérée		Relaxation labeling
H-match [CFM05]	moyenne pondérée	maxi. locale ou seuil	-
OLA [EV04]	moyenne pondérée	maxi. globale	-
oMap [ST05]	priorités entre méthodes	maxi. locale	-
OplMap [NS06]	moyenne pondérée	seuillage	consistance
QOM [ES04]	moyenne pondérée + sigmoïde	seuil + maxi. locale	-
RiMOM [TLL+06]	moyenne pondérée et sigmoïde	mini. locale du risque + seuillage	réduction de la cardinalité, consistance
SBI - Hical [IHT04]	-	seuillage	-
SCM [HYNT04]	-	maxi. locale	consistance
SF [MGMR02]	linéaire (utilise alignement d'entrée)	seuil	stable mariage
S-match/ Ctx-match [GSY04]	ensembliste par union	résolveur SAT - seuillage	

TAB. 3.3 – Comparaison des méthodes par rapport à leur composition et post-traitements

et est ainsi adaptée à de nombreux de schémas.

Au niveau des relations détectées, on remarque que la grande majorité des méthodes sont limités à l'équivalence. En effet, seule la méthode S-Match autorise plus d'expressivité dans les alignements en considérant, notamment, l'implication.

Peu de méthodes laissent un choix quant à la cardinalité des alignements qu'elles produisent. On remarque encore dans ce cas que la méthode COMA est la plus flexible sur ce point.

La table 3.3 montre, pour chaque méthode, sa composition (lorsque celle-ci combine plusieurs techniques d'alignements ou plusieurs niveaux d'information), la sélection qu'elle réalise (seuillage ou maximisation locale de l'appariement) et les post-traitements éventuellement utilisés. Nous remarquons tout d'abord qu'une grande majorité des méthodes composent les différentes techniques de manière parallèle par une combinaison statistique des résultats. La plupart utilisent à cette fin des moyennes pondérées. Quelques méthodes s'appuient également sur la fonction sigmoïde afin de favoriser le poids des meilleures valeurs et d'atténuer celui des valeurs faibles. Les méthodes basées seulement sur des approches structurelles (voir table 3.4) utilisent une composition linéaire.

On peut noter deux originalités : oMap n'utilise pas de combinaison des résultats mais est basée sur un ordre de préférence entre les différentes techniques utilisées ; S-Match réalise une combinaison ensembliste.

Les méthodes s'appuient, en général, sur une sélection des correspondances basée soit sur un seuil, soit sur la maximisation locale de l'appariement. OLA [EBB+04] utilise, quant à elle, une approche globale de maximisation de la similarité de l'alignement. RiMOM [TLL+06] utilise, entre autre, une minimisation locale du risque qui peut être assimilée au principe de maximisation locale.

Peu de méthodes proposent un post-traitement de leurs résultats. Les post-traitements les plus utilisés sont les filtres de consistance. RiMOM est la seule méthode proposant l'utilisation d'un filtre de réduction de cardinalité. GLUE [DMDH02] utilise une technique, appelée relaxation labelling, permettant, à partir de règles (générales ou spécialisées au domaine) d'affiner l'alignement produit par la méthode extensionnelle sur laquelle elle est basée. Ce post-traitement peut être vu comme une méthode d'alignement par contrainte réutilisant un alignement d'entrée.

3.6.2 Méthodes intensionnelles

La table 3.4 présente les techniques intensionnelles utilisées par chaque méthode d'alignement en distinguant les techniques terminologiques et les techniques structurelles. Parmi les méthodes présentées dans cette table, la majorité sont uniquement intensionnelles. Les méthodes GLUE, V-Doc, oMap, OplMap, QOM et RiMOM utilisent également des techniques d'alignement extensionnelles.

Au niveau des techniques syntaxiques, les distances d'édition et n-gram sont souvent utilisées. Ces mesures sont parfois appliquées également sur les chemins de noms (c.-à-d. sur la chaîne de caractères issue de la concaténation du nom de l'entité et de ceux de ses subsumantes). Ce principe peut être considéré comme une approche hybride car il prend en compte l'information structurelle sur la relation d'ordre. Quelques méthodes utilisent une distance soundex basée sur la ressemblance phonétique entre deux chaînes de caractères. RiMOM utilise une mesure de similarité statistique basée sur un corpus textuel définie par [PL02].

Au niveau linguistique, les techniques couramment utilisées s'appuient sur le thésaurus Wordnet. Certaines méthodes tirent profit des relations sémantiques (de manière stricte ou par des mesures sémantiques), d'autres exploitent seulement les synsets (ASCO, OLA). Une utilisation assez originale est celle de AUTOMS qui s'appuie sur une technique d'indexation (LSA) et des relations sémantiques pour calculer la similarité entre deux entités.

Du point de vue structurel, de nombreuses méthodes exploitent la relation d'ordre. Leur stratégie consiste à trouver de nouvelles relations à partir d'un alignement d'entrée (souvent donné par une approche syntaxique), en suivant l'idée que deux entités sont en relation si leur voisinage respectif est également en relation. Quelques méthodes proposent une approche structurelle générique, soit en confondant tous les types de relation, soit en les traitant indépendamment les uns des autres. La prise en compte de la structure interne (attribut) est souvent réalisée par une table de compatibilité entre les types de données des attributs.

Méthode	**techniques terminologiques**		**techniques structurelles**		
	syntaxique	*linguistique*	*attributs*	*relation d'ordre*	*relations transversales*
ANCHOR-PROMPT [NM01]					similarité des chemins
ASCO1 [BDKG04]	Jaro-winkler (moyenne sur les sous-chaînes des id et labels), TF.IDF (commentaires)	similarité syntaxique (Jaro-Winkler) entre synsets Wordnet, tokenization		proportions d'entités du voisinage (resp. subsumées, susbsumantes, soeurs) qui sont similaires (à partir des similarités terminologiques)	
AUTOMS (H-CONE merge [KVS06])		LSA sur thésaurus (Wordnet)			
COMA [DR02]	affixe, distance d'édition, distance n-gram, soundex (apliquées également sur chemins de noms)	synonymie (Wordnet), tokenization	table de compatibilité	moyennes sur subsumés et feuilles	
Cupid [MBR01]		basée sur thésaurus	table de compatibilité	proportion d'attributs similaires (algo. itératif)	
GLUE [DMDH02]				prise en compte des subsumants et descendants par le relaxation labelling	
GMO [HaYQW05]			Algorithme itératif de propagation de similarités		
V-Doc [QHC06]			prise en compte des entités en relation (et des attributs) pour la construction des vecteurs		
H-match [GSY04]		chemin pondéré dans Wordnet	table compatibilité		similarité contextuelle

Tab. 3.4 – Comparaison des méthodes à partir des techniques intensionnelles utilisées

Méthode	**techniques terminologiques**		**techniques structurelles**		
	syntaxique	*linguistique*	*attributs*	*relation d'ordre*	*relations transversales*
OLA [EV04]		hamming entre synsets (Wordnet)	système d'équations interdépendantes		
oMap [ST05]	égalité (chaîne entière ou racine)			moyenne des similarités des subsumant directs	
OplMap [NS06]	égalité (chaîne entière ou racine), n-gram (Jaccard) (appliquée également sur chemin de noms)			similarités binaires sur appariements des subsumants et subsumés directs	
QOM [ES04]	égalité, distance d'édition		égalité des types de données	SimSet sur subsumants, subsumés et soeurs	SimSet sur relations directes (et celles de l'ancêtre direct)
RiMOM [TLL+06]	sim. stat. de [PL02]	Lin dans Wordnet, POS Tagging, entités nommées	table de compatibilité (types de données et cardinalités)	moyenne des similarités des subsumants et des subsumés directs	moyennes des similarités du voisinage direct
SF [MGMR02]			propagation des similarités		
S-match [GSY04]	distance d'édition, distance n-gram, affixe, soundex ...	prétraitements TAL (POS-tagging), relation (stricte) dans Wordnet ($=$, $\leq$), Rada dans Wordnet			

TAB. 3.5 – Comparaison des méthodes à partir des techniques intensionnelles utilisées (suite de la table 3.4)

3.6.3 Méthodes extensionnelles

La table 3.6 donne pour chaque méthode extensionnelle recensée, le type de prétraitement qu'elle réalise, ainsi que la technique utilisée et la mesure sur laquelle s'appuie la comparaison des extensions. Les méthodes OplMap [NS06] et OMap [ST05] n'utilisent pas de mesures de comparaison étant donné que le processus de classification est implicite. En effet, avec ces méthodes, les valeurs des prédictions sont combinées (par une moyenne pondérée).

Au niveau des prétraitements utilisés, les méthodes recensées utilisent soit la classification des instances par bayésien naïf (ou une amélioration), soit de la réindexation vectorielle. QOM n'utilise pas prétraitement étant donnée que la comparaison des extensions repose sur l'agrégation, par la mesure SimSet [ES04], des valeurs de similarités inviduelles.

Conclusion

Nous avons présenté dans ce chapitre différents aspects des méthodes d'alignement. Nous avons tout d'abord défini la notion de méthode d'alignement, puis nous les avons étudiées selon trois axes principaux.

Le premier axe d'étude concerne les entrées et sorties des méthodes d'alignement. En effet, une méthode prend, en entrée, des représentations structurées, qui peuvent être de différents formats : hiérarchies textuelles, schémas objets, relationnels, ontologies (OWL/RDFS). En sortie, les méthodes se distinguent quant à la nature des alignements qu'elle produisent. Ces alignements peuvent être distingués selon leur cardinalité (fonctionnels, injectifs, ou de cardinalité quelconque) ou encore selon les relations qu'ils contiennent (relation d'équivalence uniquement ou relation d'implication).

Le deuxième axe est celui de la composition globale des méthodes d'alignement. Nous avons distingué la composition parallèle (qui peut être ensembliste ou statistique) et la composition linéaire.

Le troisième et dernier axe s'intéresse aux techniques locales de comparaisons utilisées. Ces techniques peuvent être réparties en deux grandes familles : les techniques intensionnelles et les techniques extensionnelles. Concernant, les techniques intensionnelles, nous sommes basés sur les classifications de [RB01] et [SE05], qui distinguent les approches terminologiques (structurelles et linguistiques) des approches structurelles (internes et externes). Pour les techniques extensionnelles, nous sommes partis du constat que ces méthodes réalisent un pré-traitement visant à redéfinir les hiérarchies sur une extension commune. De ce fait, nous avons distingué les approches fonctionnant par augmentation, utilisant la classification supervisée, des approches de réindexation de l'extension.

Finalement, à partir de ces trois axes d'étude, nous avons proposé une comparaison synthétique d'une vingtaine de méthodes proposées dans la littérature. Il ressort de cette comparaison, qu'une grande majorité des méthodes sont basées sur des combinaisons de mesures de similarité. Par conséquent, elles permettent de détecter seulement des relations d'équivalence entre entités. Les seules méthodes considérant la relation d'implication sont uniquement basées sur une reconnaissance stricte de cette relation à partir d'une base de données lexicales.

Méthode	**Type de prétraitement**	**Technique utilisée**	**Comparaison des extensions**
CAIMAN [LG01]	Réindexation vectorielle	TF.IDF - Rocchio	cosinus
GLUE [DMDH02]	Classification	bayésien naïf sur contenu et chemin de noms	Jaccard
V-Doc [QHC06]	Réindexation vectorielle	TF.IDF (prise en compte des labels, commentaire, etc.)	cosinus
oMap [ST05]	Classification	bayésien naïf	aucun
OplMap [NS06]	Classification	bayésien naïf + k-plus-proches voisins	aucun
QOM [ES04]		aucun	SimSet sur les instances directes et celles des subsumés directs
RiMOM [TLL+06]	Classification	bayésien naïf (appliquée également sur données prétraitées par POS et entités nommées)	?
SBI - Hical [IHT04]	Classification	Adaptation bayésien naïf aux hiérarchies	kappa
SCM [HYNT04]	Réindexation vectorielle	Kullback-Leibler avec espace vectoriel non euclidien	cosinus

Tab. 3.6 – Comparaison des méthodes à partir des techniques extensionnelles utilisées

4 La méthode AROMA

Sommaire

Introduction

Ce chapitre présente la méthode AROMA (Association Rule Ontology Matching Approach). Contrairement aux méthodes présentées dans le chapitre précédent, AROMA a l'originalité de détecter des relations d'implication entre entités issues de deux hiérarchies en s'appuyant sur les données textuelles contenues dans l'extension (instances) et dans les annotations (nom, commentaire, etc.).

Nous introduisons, dans un premier temps, les principes généraux et la composition de la méthode AROMA. Ensuite, nous présentons deux approches de pré-traitement des hiérarchies d'entrée : une dédiée aux hiérarchies textuelles et une autre destinée aux ontologies RDFS/OWL. Finalement, nous présentons l'extraction de l'alignement en incluant, d'une part, les descriptions des filtres de post-traitement et, d'autre part, la présentation d'une méthode intensionnelle et syntaxique d'enrichissement de l'alignement.

4.1 Principes généraux et composition d'AROMA

La méthode AROMA permet la découverte d'un alignement entre deux hiérarchies contextualisées $\mathcal{H}_1 = (C_1, \leq, \mathcal{A}_1, O_1, \sigma_1)$ et $\mathcal{H}_2 = (C_2, \leq, \mathcal{A}_2, O_2, \sigma_2)$.

Notre méthode d'alignement suit le schéma classique de l'ECD composé d'une phase de pré-traitements préparant les hiérarchies au processus fouille de règles qui sera ensuite suivi d'une dernière étape de post-traitements permettant de finaliser l'alignement. La composition globale de notre méthode AROMA est schématisée sur la figure 4.1.

L'étape de pré-traitement permet de préparer les deux hiérarchies $\mathcal{H}_1$ et $\mathcal{H}_2$ en redéfinissant leurs relations d'association respectives σ_1 et σ_2 sur une extension commune. Pour cela, on peut utiliser une des approches de classification présentées dans la section 3.5 ou une des deux approches de redéfinition de l'extension que nous proposons dans la section suivante. Dans les cas où les deux hiérarchies à comparer sont définies sur un ensemble de termes ou de textes communs, cette étape de pré-traitement n'est alors pas nécessaire. Cependant, cela est rarement le cas dans la pratique.

Pour le processus de fouille de règles, nous adoptons une approche asymétrique permettant de découvrir un ensemble d'éléments de correspondance implicatifs entre les entités d'une première hiérarchie $\mathcal{H}'_1$ vers les entités d'une seconde hiérarchie $\mathcal{H}'_2$. A cause de cette nature asymétrique, l'algorithme devra être exécuté deux fois pour obtenir l'ensemble des correspondances entre les deux hiérarchies : une première fois pour découvrir les implications issues des entités de $\mathcal{H}'_1$ vers les entités de $\mathcal{H}'_2$, et une deuxième afin de découvrir les implications issues des entités de $\mathcal{H}'_2$ vers les entités de $\mathcal{H}'_1$. A l'issue de ce processus de fouille, les deux alignements asymétriques sont fusionnés pour donner un alignement implicatif.

Ensuite AROMA s'appuie sur une phase de post-traitement. Cette phase comporte une première étape permettant de déduire les éléments de correspondance en relation d'équivalence. A l'issue de cette étape, nous obtenons un alignement symétrique de cardinalité $0, n - 0, n$. Cependant, cet alignement peut contenir des inconsistances. Nous proposons un filtre permettant de détecter et de supprimer les situations où des éléments de correspondance sont en incohérence. Finalement, afin d'obtenir un alignement de cardinalité restreinte et/ou asymétrique, nous proposons également des filtres permettant respectivement d'extraire la composante asymétrique et de rendre un alignement fonctionnel.

Dans un second temps, il est possible d'enrichir l'alignement fourni par le premier algorithme. En effet, cette seconde étape est utile lorsque les données extensionnelles ne sont pas assez conséquentes et que certaines correspondances n'ont pas pu être extraites. Nous proposons alors un algorithme s'appuyant sur l'intension des hiérarchies. Ce deuxième algorithme, de nature symétrique, permet de sélectionner des correspondances en utilisant une approche terminologique syntaxique (similarités sur chaînes de caractères). Afin d'améliorer la pertinence de la recherche, la méthode proposée s'appuie sur l'alignement préalablement calculé ainsi que sur les relations d'ordre des hiérarchies.

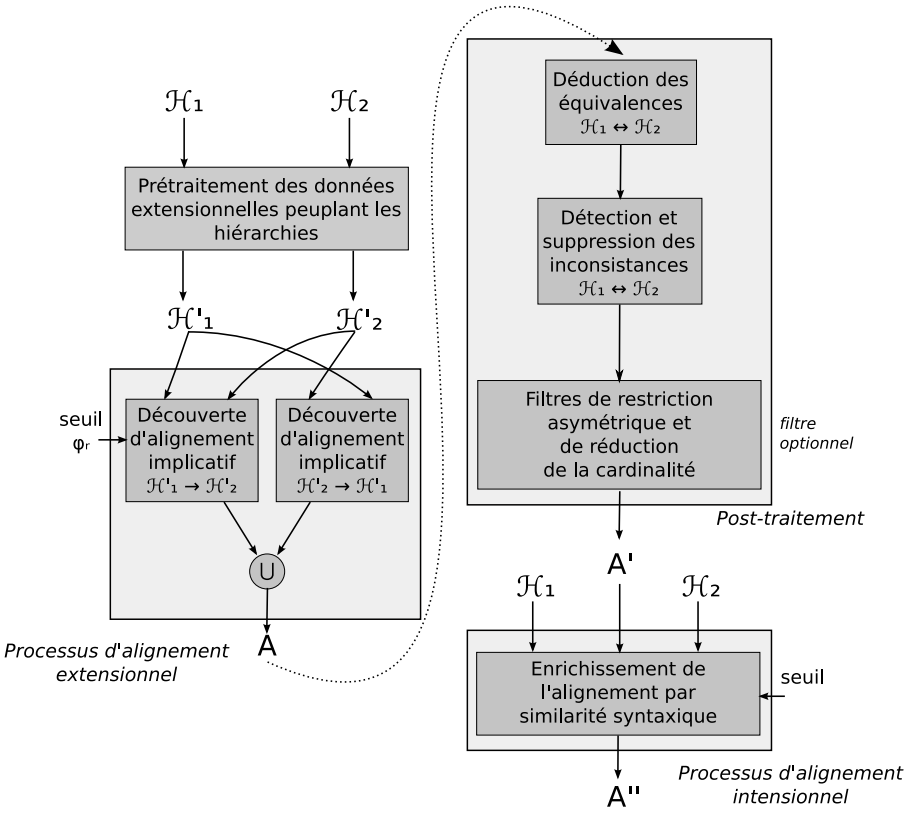


FIG. 4.1 – Schéma de composition d'AROMA

Finalement, un dernier filtre permet d'éliminer les éventuels éléments de correspondance redondants afin d'obtenir une alignement minimal.

4.2 Réindexation de hiérarchies

L'alignement de hiérarchies s'appuyant sur leurs données extensionnelles nécessite qu'elles partagent une intersection conséquente. Cependant, il est rare que les entités issues de deux hiérarchies soient associées aux mêmes ensembles d'objets. Afin de pallier ce problème, nous proposons deux méthodes de prétraitement des hiérarchies. Ces méthodes permettent de redéfinir les relations d'association des entités sur des nouveaux ensembles constitués de descripteurs issus des objets et des annotations associées aux entités.

La première méthode est dédiée aux hiérarchies textuelles, où les objets associés aux entités sont des documents textuels. Cette méthode permet d'extraire les termes contenus dans les documents, de les sélectionner et de les associer aux entités. Cette méthode est, entre autre, applicable aux répertoires Web, catalogues de boutiques en ligne, forums de discussions, etc.

La deuxième méthode est conçue pour représenter les ontologies au format

RDFS/OWL en deux hiérarchies : l'une représentant la hiérarchie des classes, et l'autre, la hiérarchie des propriétés. La nouvelle relation d'association sera constituée de données issues des annotations et des instances. Ces données peuvent être des termes (extraits des annotations) ou des valeurs d'attributs (extraites des instances).

Ces étapes de pré-traitement permettent de redéfinir les extensions des hiérarchies sur des ensembles constitués de données moins complexes que les objets initialement associés. En effet, des ensembles de termes (ou autres données de type simples) sont d'une part, plus facilement comparables que des ensembles textes entiers ou des structures complexes telles que des classes ou des propriétés RDFS/OWL et ont, d'autre part, de plus grandes chances d'être en partie partagés par les deux structures. Ainsi les hiérarchies sont susceptibles d'avoir une intersection plus volumineuse qu'initialement. A l'issue de cette première étape, l'extraction de règles d'association entre hiérarchies sera, de ce fait, plus aisée.

4.2.1 Pré-traitement d'une hiérarchie textuelle

Cette étape de pré-traitement, destinée à des hiérarchies peuplées par objets de type documents textuels, permet de les repeupler sur un ensemble de termes extraits à partir des documents. Prenons, par exemple, une entité c de la hiérarchie et un terme t apparaissant dans au moins un des documents associés à la hiérarchie. L'intuition sous-jacente à notre approche de pré-traitement est que le terme t sera représentatif de l'entité c si les documents dans lesquels le terme t apparaît tendent à être associés (par la relation d'association) à l'entité c.

Notre méthode d'association aux entités de leurs termes représentatifs est divisée en deux étapes :

1. L'indexation terminologique des documents permettant de représenter chaque document par les termes qu'il contient.
2. La sélection pour chaque entité d'un ensemble de termes dits représentatifs.

Formalisation

Etant donnée une hiérarchie $\mathcal{H} = (C, \leq, \mathcal{A}, D, \sigma)$, où l'ensemble D est constitué de documents textuels. Ces documents peuvent être, par exemple, des descriptifs de produits d'un catalogue hiérarchique, des descriptifs (ou pages d'accueil) de sites Web indexés par un répertoire Web, des documents indexés par un thésaurus, etc. Chaque document $d \in D$ peut être décrit par l'ensemble des termes $T_d = \{t_1, ...t_n\}$ qu'il contient. L'ensemble des termes contenus dans les documents de D est noté T_0.

A l'issue de l'étape de pré-traitement, chaque entité $x \in C$ sera associée à son ensemble de termes représentatifs $\sigma'(x)$. Ainsi la hiérarchie initiale $\mathcal{H}$, munie de sa nouvelle extension, sera définie par le quintuplet $\mathcal{H}' = (C, \leq, \mathcal{A}, T, \sigma')$. L'ensemble T ($T \subseteq T_0$) représentera l'ensemble des termes associés aux entités par la relation σ'. Afin de conserver la propriété d'isomorphisme entre les ensembles

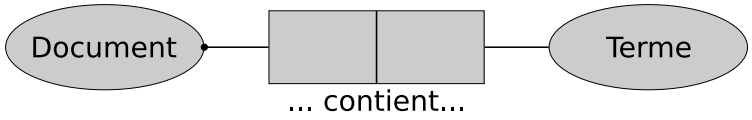


FIG. 4.2 – Diagramme NIAM de la relation entre les termes et les documents

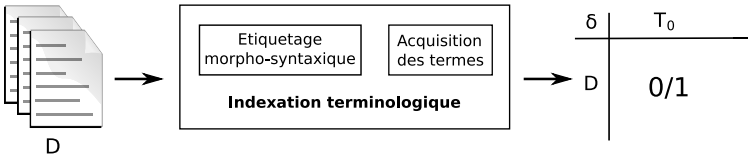


FIG. 4.3 – Indexation terminologique

ordonnés $(C, \leq)$ et $(2^O, \subseteq)$, un terme associé à une entité x devra également être associé à toutes ses entités majorantes : si $x \leq x'$ alors $\sigma'(x) \subseteq \sigma'(x')$.

1ère étape : Indexation terminologique

Dans le domaine de la recherche d'information, l'indexation automatique permet d'assigner, à un texte, un ensemble de descripteurs qualifiant le document. On recense deux types d'indexation : (1) l'indexation contrôlée qui indexe les textes par rapport à un référentiel prédéterminé (par exemple, un thésaurus) ; (2) l'indexation libre qui n'utilise, quant à elle, pas de référentiel. Nous nous sommes intéressés à ce dernier type d'indexation où les descripteurs potentiels sont issus des textes. Ces méthodes d'indexation reposent sur une première phase consistant à acquérir les termes à partir des textes.

Le processus d'acquisition des termes à partir de données textuelles consiste à modéliser chaque texte par un ensemble de termes qu'il contient. A l'instar de [Che04], nous nous intéressons seulement à l'apparition d'un terme dans un texte, nous ne prenons pas en compte le nombre d'occurrences de ce terme dans le texte. Cette modélisation est représentée par le diagramme NIAM [1] figure 4.2. Un texte possède un ensemble de termes et un terme caractérise un ensemble de textes. Cette modélisation du texte est réalisée en deux temps. Tout d'abord, une première étape de pré-traitement permet d'annoter les textes afin de pouvoir leur appliquer, dans un second temps, un extracteur de termes.

Cette première étape d'indexation, illustrée figure 4.3, permet de représenter chaque texte par les termes qu'il contient. A partir de cette représentation, nous définissons la relation d'indexation $\delta : T_0 \times D$ qui relie les termes aux documents ($\delta(t)$ dénote ainsi l'ensemble des documents dans lesquels le terme t apparaît). L'ensemble des termes extraits est appelé T_0. La relation δ est représentée par une matrice booléenne de présence (1)/absence (0) d'un terme $t \in T_0$ dans un texte $d \in D$.

[1] Natural language Information Analysis Method

Etiquetage morpho-syntaxique et lemmatisation. L'étiquetage morpho-syntaxique (Part-Of-Speech (POS) tagging [Chu88]), consiste à associer une étiquette grammaticale à chaque mot d'un texte en fonction du contexte. Une étiquette est composée de la forme grammaticale du mot (nom, adjectif, verbe, adverbe, préposition, article, etc.) ainsi que d'informations supplémentaires telles que le genre (masculin, féminin), le nombre (singulier, pluriel), ou encore du temps et de la personne dans le cas des verbes. Les étiqueteurs sont basés sur des modèles probabilistes permettant de prédire l'étiquette d'un mot en fonction des étiquettes du ou des mots précédents. Les étiqueteurs sont entraînés sur des corpus d'apprentissage et donnent généralement de très bons résultats (précision de l'ordre de 99,5% sur l'anglais).

Après la phase d'étiquetage, le processus de lemmatisation permet d'associer à chaque mot du texte sa forme canonique. Par exemple, le verbe « marcher » peut apparaître dans une phrase sous les formes « marchant », « marché », « marchais », etc. La lemmatisation permettra à partir des connaissances grammaticales extraites par l'étiqueteur morpho-syntaxique et d'un dictionnaire, de reconnaître l'occurrence du verbe et sa forme canonique « marcher ». Pour un verbe, sa forme canonique sera l'infinitif, pour les noms, le singulier et les autres mots (adjectifs, pronoms, articles) le masculin singulier.

Il existe aussi une opération similaire utilisée en recherche d'information appelée racinisation (stemming, en anglais). La différence avec la lemmatisation est qu'une opération de racinisation n'utilise pas de connaissances sur le contexte (et est donc utilisable directement, sans étiquetage morpho-syntaxique). Cette opération est plus simple à implémenter et plus rapide sur l'analyse d'un texte mais elle ne permet pas de discriminer certains mots ayant une syntaxe similaire mais un sens différent. L'exemple précédent sur le mot « marché » est un exemple de mots ayant deux sens (le verbe marcher et le mot marché désignant le lieu de vente) qui sont discriminables selon leur contexte grammatical.

Acquisition des termes. Après la première phase de prétraitement, les outils de TALN permettent de représenter un texte par un ensemble de termes qu'il contient. Nous retenons la définition suivante d'un terme :

Définition 4.1 *Un terme est un syntagme nominal, c'est à dire un ensemble d'un ou plusieurs mots dont le noyau (élément central du syntagme) est un nom.*

Nous avons choisi de ne pas extraire seulement des termes simples (c.-à-d. composés d'un seul mot) mais également les syntagmes nominaux car ils sont plus informatifs et présentent ainsi l'avantage de ne pas avoir d'ambiguïté lexicale par rapport aux termes simples. Par exemple, les termes « avocat d'affaire » ou « avocat du barreau » sont des termes qui présentent l'avantage d'enlever l'ambiguïté que l'on aurait eu avec le terme simple « avocat » qui peut désigner à la fois un métier ou un légume.

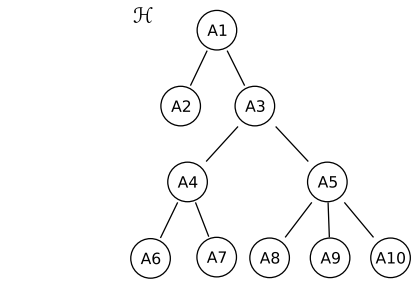
On recense deux familles d'approches de découverte automatique de termes en corpus : les approches probabilistes et les approches structurelles [Dai94]. Les méthodes probabilistes s'appuient, soit sur l'évaluation des cooccurrences pour extraire des termes, soit sur le comptage des segments répétés [LS88]. Les méthodes structurelles repèrent les syntagmes nominaux en corpus par schémas

positifs, ou encore par leur bornes. Les méthodes structurelles par schémas positifs utilisent des patrons morpho-syntaxiques (de type « NOM PREPOSITION NOM », « NOM ADJECTIF », « ADJECTIF NOM », etc.), pour le repérage des syntagmes. Les méthodes structurelles d'acquisition de termes par leur bornes [Bou94] reposent sur la constatation que certains constituants d'une phrase ne peuvent pas faire partie d'un syntagme nominal. Ces constituants, appelés bornes, sont les adverbes, les verbes conjugués, la ponctuation, les pronoms. Ces méthodes permettent d'extraire des groupes nominaux dits maximaux. Finalement, des méthodes dites mixtes combinent à la fois l'analyse probabiliste et l'analyse structurelle [Sma93], [Dai94].

Dans le cadre des méthodes structurelles, une seconde étape d'analyse des variations des termes peut être menée afin de repérer et de regrouper toutes les variantes d'un terme. Dans [Dai03], 7 types de variations sont recensés :

- **Graphiques** : Ce type de variantes concerne le changement de casse ou la présence de tiret dans une structure « NOM NOM ». Par exemple le terme « mot-clé » peut aussi s'écrire également « mot clé ».
- **Flexionnelles** : Ce sont des variantes orthographiques d'un des noms constituant le terme, comme par exemple l'ajout d'un « s » au pluriel.
- **Syntaxiques faibles** : Ces variations concernent le changement de la préposition, l'ajout ou la suppression d'une préposition et/ou d'un déterminant dans le terme ou encore le changement de la position de l'adjectif (passage d'épithète en position d'attribut).
- **Syntaxiques** : Ces variations sont obtenues par ajout d'un modifieur (adjectif, adverbe, ou encore les spécifieurs de type nominal) dans le terme (exemple : le terme « crème fouettée » devient « crème fraîche fouettée ») ou par variation de coordination (par exemple : les termes « vin sec » et « vin blanc » peuvent être combinés en « vin blanc et sec »).
- **Morpho-syntaxiques** : Ces variations regroupent les termes dont un des constituants a subi une modification de sa structure par dérivation lexicale : adjonction d'un affixe ou utilisation d'un adjectif relationnel. Par exemple, le terme « distance entre classes » devient par variation morphologique « distance inter-classes ». Le terme « variable de Poisson » devient « variable poissonienne » en utilisant l'adjectif relationnel.
- **Paradigmatiques** : Ce sont les variations basées sur le principe de substitution de la linguistique distributionnelle (Harris). Elle consiste par remplacer (au moins) un constituant du terme par son synonyme. Exemple : « alignement de hiérarchies » et « alignement de taxonomies » (si l'on considère que les mots taxonomies et hiérarchies sont synonymes).
- **Anaphoriques** : Ce sont les variations mettant en jeu des sigles ou utilisant seulement le constituant de tête du terme pour faire référence au terme complet. Par exemple, dans un contexte donné, le terme « variable » peut être employé pour désigner le terme « variable de Poisson ». l'abréviation « v.a. » est également une variation anaphorique du terme « variable aléatoire ».

Afin d'extraire les termes des données textuelles, nous nous sommes appuyés sur une approche de détection par patrons syntaxiques, puis de regroupement par analyse des variations. Pour cela, nous avons utilisé ACABIT [Dai94], une méthode mixte qui permet également le classement des termes par calculs de différents indices. Notre objectif étant d'indexer les documents par les termes



D

σ	A2	A1	A6	A4	A7	A3	A8	A5	A9	A10
d1	1	1	0	0	0	0	0	0	0	0
d2	1	1	0	0	0	0	0	0	0	0
d3	0	1	0	0	0	0	0	0	0	0
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
dn	0	1	0	0	0	1	0	1	1	0

C

δ	t1	t2	t3	...	...	tm
d1	1	0	0	.	.	0
d2	1	0	0	.	.	0
d3	0	0	0	.	.	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
dn	1	1	0	.	.	0

T_0

FIG. 4.4 – Indexations σ et δ

qu'ils contiennent, nous n'utilisons pas l'ordonnancement des termes proposé par ACABIT. Les termes pris en compte par ACABIT sont des termes binaires c.-à-d. composés de deux mots significatifs. En effet, cette méthode utilise les patrons syntaxiques suivants : « NOM ADJECTIF », « NOM (PREP (DET)) NOM », et « NOM à VERBE INFINITIF ». Les parenthèses dans les patrons représentent des séquences optionnelles.

2ème étape : Sélection des termes représentatifs des entités

A partir des termes préalablement extraits des textes associés à la hiérarchie, cette seconde étape consiste à sélectionner et à associer un ensemble de termes pour chaque entité de la hiérarchie. Ces termes doivent être représentatifs du vocabulaire utilisé dans les documents associés initialement à l'entité c.

Nous disposons de deux relations d'indexation booléennes σ et δ, représentées sur la figure 4.4. La première, qui fait partie de la hiérarchie, associe les textes aux entités de la hiérarchie, la seconde, que nous avons construit, indexe les textes par les termes qu'ils contiennent. A partir de ces deux indexations, nous voulons associer à chaque entité, un ensemble de termes représentatifs du vocabulaire utilisé dans les documents indexés à cette entité.

Plusieurs approches sont possibles pour réaliser les sélections des termes représentatifs d'une entité. La première est une approche symétrique visant à évaluer la tendance $t \leftrightarrow c$. Cette tendance permettra d'associer un terme t à une entité c si l'on observe relativement souvent des documents contenant le terme t et associés à c. Les deux autres sont asymétriques. On peut tout d'abord étudier la tendance $c \rightarrow t$, c.-à-d. associer un terme t à l'entité c si les documents associés à c ont tendance à contenir également le terme t. La dernière approche s'intéresse

à la tendance inverse $t \rightarrow c$, reposant sur le principe d'associer le terme t à l'entité c si les documents dans lesquels apparaît le terme t ont tendance à être associés à l'entité c.

Nous avons choisi cette dernière tendance $t \rightarrow c$. Ce choix est justifié par l'étude comparative de la distribution des nombres de documents contenant un terme donné, et celle des nombres de documents associés à une entité. Les figures 4.5, 4.6 et 4.7 montrent respectivement, pour plusieurs jeux de tests, ces deux distributions. Celles de droite donnent la probabilité (en ordonnée) qu'une entité soit associée à x documents (en abscisse). Celles de gauche donnent la probabilité (en ordonnée) qu'un terme (apparaissant plus d'une fois dans le corpus), soit contenu dans x documents (en abscisse).

Ces distributions montrent clairement une asymétrie entre les deux phénomènes : un terme apparaît généralement dans relativement peu de documents alors qu'une entité est associée à beaucoup plus de documents. Cette asymétrie, illustrée par le diagramme de Venn figure 4.8, montre qu'un terme devra être associé à une entité si l'ensemble de documents dans lesquels le terme apparaît a tendance à être inclus dans l'ensemble des documents associés à l'entité considérée.

Afin de prendre en compte cette asymétrie, le modèle des règles d'association [AIS93] semble adéquat car il permet de découvrir des tendances orientées (et donc asymétriques) entre ensembles d'attributs. Notamment, nous nous limitons à l'étude des règles du type $t \rightarrow c$ dont la sémantique (pour une règle valide) est la suivante : « Si un document contient le terme t alors ce document a tendance à être associé à l'entité c ». Afin de vérifier la validité de cette tendance, nous évaluons, pour chaque couple (t,c), la qualité la règle d'association $t \rightarrow c$ sur l'ensemble des documents.

Nous avons choisi d'évaluer la qualité de la règle d'association $t \rightarrow c$ par la mesure d'Intensité d'Implication [Gra96], permettant d'estimer la probabilité que le nombre de contre-exemples de la règle $t \rightarrow c$ (c.-à-d. le nombre de documents contenant le terme t et qui ne sont pas associés à l'entité c) soit plus petit que celui attendu sous hypothèse d'indépendance des parties X et Y, tirées de manière aléatoire et de même cardinalité que $\delta(t)$ et $\sigma(c)$. La modélisation de cette mesure, schématisée figure 4.9, est donnée par :

$$\varphi(t \rightarrow c) = 1 - Pr\left(N_{t\overline{c}} \leq n_{t\overline{c}}\right)$$

- $n_{t\overline{c}} = |\delta(t) - \sigma(c)|$ est le nombre de contre-exemples observés c.-à-d. les documents qui contiennent le terme t et qui ne sont pas associés à l'entité c.
- $N_{t\overline{c}}$ est le nombre aléatoire de contre-exemples obtenu sous hypothèse d'indépendance.

Sous hypothèse d'indépendance, la probabilité de tirer un document contenant le terme t qui n'est pas associé à c est $P(t\overline{c}) = P(t) \times P(\overline{c}) = \frac{|\delta(t)| \times |D - \sigma(c)|}{|D|^2}$. Ainsi, l'espérance de la variable $N_{t\overline{c}}$ est égale à $\lambda = \frac{|\delta(t)| \times |D - \sigma(c)|}{|D|} = \frac{n_t \times (N - n_c)}{N}$.

Définition 4.2 *Un terme t sera **représentatif** pour une entité c, relativement au corpus D, si $\varphi(t \rightarrow c)$ est supérieure à un seuil fixé φ_t.*

Le choix du seuil φ_t est laissé à l'utilisateur. Ce seuil compris entre 0 et 1,

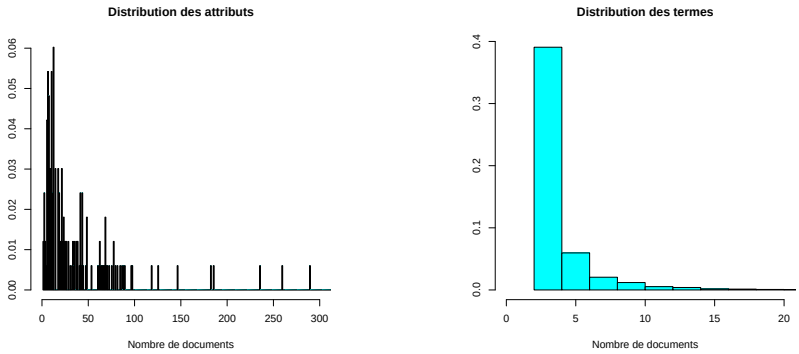


FIG. 4.5 – Distributions associées au catalogue de cours Cornell

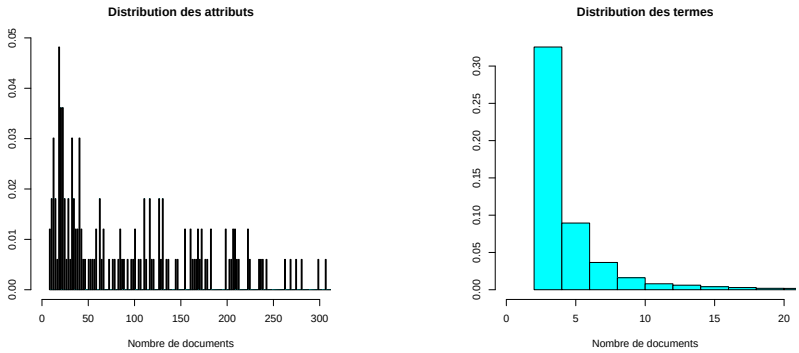


FIG. 4.6 – Distributions associées au catalogue de cours Washington

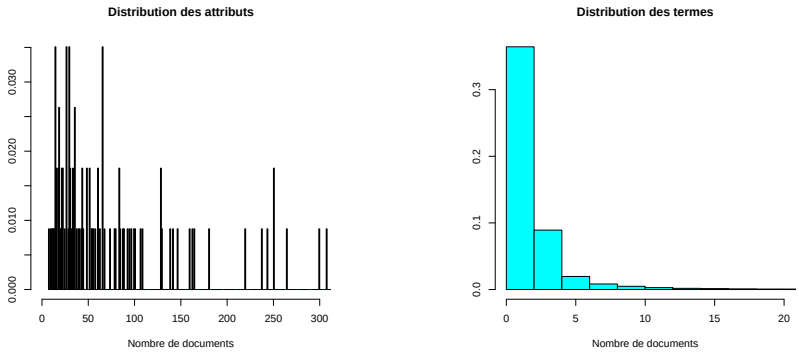


FIG. 4.7 – Distributions associées au répertoire Web Yahoo Finance

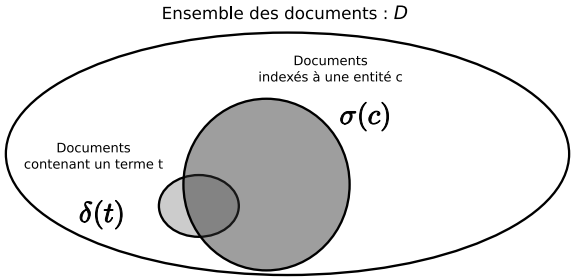


FIG. 4.8 – Diagramme de Venn représentant les ensembles de documents contenant un terme t et ceux indexés à une entité c

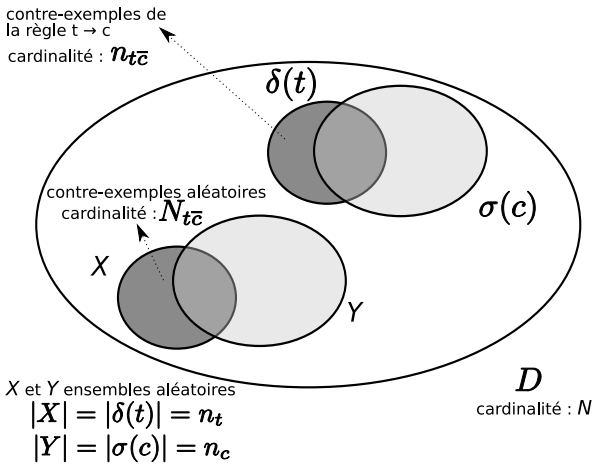


FIG. 4.9 – Intensité d'implication d'une règle $t \rightarrow c$

doit être, dans la pratique, supérieur à $0,5$. En effet, la valeur de $0,5$ est prise par l'Intensité d'Implication lorsque la règle satisfait l'indépendance entre les documents contenant un terme t et ceux associés à une entité c.

A partir de la définition 4.2, l'ensemble des termes représentatifs de l'entité c est $\sigma'(c)$:

$$\sigma_0(c) = \{t \in T_0 | \varphi(t \rightarrow c) \geq \varphi_t\}$$

Afin d'assurer la propriété d'isomorphisme entre les ensembles ordonnés $(C, \leq)$ et $(2^T, \subseteq)$, l'ensemble des termes associés à une entité c est constitué des termes représentatifs de c ainsi que des termes représentatifs des entités qu'elle majore. L'ensemble des termes $\sigma'(c)$ associés à une entité c est alors défini par :

$$\sigma'(c) = \sigma_0(c) \cup \left\{ \bigcup_{c' < c} \sigma_0(c') \right\}$$

A l'issue de cette étape de pré-traitement, la hiérarchie $\mathcal{H}$ initialement définie sur un ensemble de documents sera redéfinie sur un ensemble de termes par $\mathcal{H}' = (C, \leq, T, \sigma')$ où T désigne l'ensemble des termes significatifs sélectionnés.

4.2.2 Ontologies RDFS/OWL

Les ontologies RDFS/OWL sont définies comme des vocabulaires structurés décrivant un ensemble de concepts (classes) ainsi leurs relations. Le langage OWL permet de définir des classes et de les organiser en taxonomies en utilisant la relation de spécialisation. Les classes peuvent être décrites intensionnellement en utilisant le prédicat RDFS *subClassOf* ou grâce aux prédicats OWL *intersectionOf*, *unionOf* et *complementOf*. Ce langage permet également de définir les classes de manière extensionnelle comme des énumérations d'objets avec la propriété *oneOf*. Dans le but d'établir les relations entre les classes, OWL permet de déclarer des propriétés et de les organiser elles aussi en taxonomies. Une propriété peut être transitive, symétrique ou fonctionnelle. Une propriété possède un domaine et un codomaine. Le type du domaine d'une propriété est une classe tandis que celui de son codomaine est soit une classe, soit un type de donnée élémentaire tel qu'un entier ou une chaîne de caractères. OWL permet d'instancier des classes (individus) et d'assigner des valeurs à chacune de leurs propriétés. Le langage OWL fournit également des prédicats tels que l'équivalence ou l'incompatibilité entre classes. Chaque classe, propriété ou individu peut être annoté à l'aide des primitives *owl :versionInfo*, *rdfs :label*, *rdfs :comment*, *rdfs :seeAlso*, et *rdfs :isDefinedBy*.

Formalisation des hiérarchies d'une ontologie RDFS/OWL

Les ontologies décrites en RDFS et OWL possèdent deux structures hiérarchiques : la hiérarchie des classes et celle des propriétés. Pour simplifier, nous représentons une ontologie RDFS/OWL par un couple de hiérarchies :

$$\mathcal{O} = (\mathcal{H}_c, \mathcal{H}_p)$$

La hiérarchie des classes est représentée par le quintuplet :

$$\mathcal{H}_c = (C, \leq, \mathcal{A}, I_c, \sigma_c)$$

- C désigne l'ensemble des classes.
- $\leq$ est la relation de subsomption entre les classes.
- I_c dénote l'ensemble des instances des classes.
- σ_c est la relation associant à chaque classe, son ensemble d'instances.

La hiérarchie des propriétés est représentée par le quintuplet :

$$\mathcal{H}_p = (P, \leq_p, \mathcal{A}, I_p, \sigma_p)$$

- P désigne l'ensemble des propriétés.
- $\leq$ est la relation de subsomption entre les propriétés.
- I_p dénote l'ensemble des instances des propriétés.
- σ_p est la relation associant à chaque propriété, son ensemble d'instances.

La fonction $\mathcal{A}$, commune aux deux hiérarchies, regroupe les fonctions d'annotations des langages RDFS et OWL. Ces fonctions permettent d'associer à chaque classe, propriété, instance (de classe ou de propriété), des annotations telles qu'un identifiant, un ensemble de labels et des commentaires. Les correspondances dans le langage RDFS des fonctions d'annotation contenues dans $\mathcal{A}$ sont :

- $\mathcal{A}_{id}$: identifiant *rdfs :id*.
- $\mathcal{A}_{label}$: labels *rdfs :label*. On peut associer un label par langue.
- $\mathcal{A}_{com}$: commentaires *rdfs :comment*. On peut associer un commentaire par langue.

Formalisation de la réindexation

Le pré-traitement des hiérarchies va consister, pour chaque entité (classes et propriétés), à lui réassocier un ensemble d'objets contenant les termes extraits à partir de ses annotations propres et celles de ses instances. Nous ajouterons également à cet ensemble d'objets, les valeurs prises par les attributs contenus dans ses instances. Ces dernières valeurs peuvent être soit des termes dans le cas d'attribut de type chaîne de caractères, soit des données autres (telles que des entiers, dates, etc.) dans les autres cas.

Ainsi chaque hiérarchie $\mathcal{H}_c$ et $\mathcal{H}_p$ sera repeuplée sur un ensemble de termes et de données, noté T_c et T_p. Ces hiérarchies pourront être alors réécrites par :

$$\mathcal{H}'_c = (C, \leq, \sigma'_c, T_c)$$

$$\mathcal{H}'_p = (P, \leq, \sigma'_p, T_p)$$

T_c et T_p représentent les ensembles des termes et des données associés aux classes et aux propriétés. Chaque classe (resp. propriété) sera associée à un sous-ensemble de T_c (resp. T_p) par la relation d'association σ'_c (resp. σ'_p).

Avant de présenter les définitions des ensembles de données des classes et des propriétés, nous introduisons les fonctions suivantes qui permettent d'accéder aux différentes valeurs des entités :

- $Termes : 2^{Texte} \longrightarrow 2^{Texte}$: Cette fonction d'acquisition terminologique retourne l'ensemble des termes binaires et simples contenus dans un ensemble de textes fourni en entrée.
- $\mathcal{V} : I_c \longrightarrow 2^{Texte}$: Cette fonction retourne l'ensemble des valeurs prises par les propriétés de type *owl :datatypeProperty* pour l'instance donnée.
- $\mathcal{O} : I_p \longrightarrow I_c \cup 2^{Texte}$: Cette fonction retourne l'objet d'une instance de propriété. L'objet peut être soit une instance (dans le cas d'une propriété de type *owl :ObjectProperty*), soit une valeur littérale (dans le cas d'une propriété de type *owl :DatatypeProperty*).

La notation 2^{Texte} représente l'ensemble des ensembles de chaînes de caractères.

Exemple. Intéressons-nous à l'extrait d'ontologie OWL présenté ci-dessous. Sur cet extrait, la classe « Muscadet » est définie comme une sous-classe de « vin blanc ». Elle est spécialisée par restriction sur les propriétés « produitDans » et « Cépages », qui représentent respectivement la région de production du vin et les cépages dont il est constitué. La classe « Muscadet » est ainsi définie comme un « vin blanc » produit dans la région de Nantes et constitué du cépage Melon de Bourgogne. L'extrait OWL contient aussi la description d'une instance de la classe « Muscadet » qui est une bouteille de muscadet d'appelation « Muscadet sèvre et Maine sur Lie » de millésime 1998 dont le producteur est « Domaine de Truc ».

```
<owl:Class rdf:ID="Muscadet">
  <rdfs:subClassOf rdf:resource="#VinBlanc" />
  <rdfs:label>
    Muscadet
  </rdfs:label>
  <!-- Commentaire extrait de Wikipedia -->
  <rdfs:comment>
    Le Muscadet est un type de vin blanc sec français issu de la vallée
    de la Loire au sud de Nantes. Les vins du Muscadet sont vinifiés
    en blanc sec à partir d'un cépage unique, le melon (on le retrouve
    aussi sous les dénominations de melon de Bourgogne, gamay de
    Bourgogne ou melon musqué). La dénomination sur lie (ou sur lies) peut
    être ajoutée à l'appellation. Dans ce cas, les vins doivent avoir
    passé un seul hiver en fûts ou en cuves et se trouver encore sur
    leurs lies de fermentation au moment de la mise en bouteille.

  </rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#ProduitDans" />
      <owl:hasValue rdf:resource="#RegionNantes" />
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#Cepages" />
      <owl:hasValue>Melon de Bourgogne</owl:hasValue>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>

<Muscadet rdf:ID="MaBouteilleDeMuscadet">
  <Produitpar rdf:resource="#DomaineDeTruc"/>
  <Millesime>1998</Millesime>
  <Appellation>
    Muscadet Sèvre-et-Maine sur Lie
  </Appellation>
  ...
</Muscadet>

<Producteur rdf:ID="DomaineDeTruc">
```

```
  <Nom>Domaine de Truc</Nom>
  <Ville>LA HAYE FOUASSIERE</Ville>
</Producteur>
```

Cette classe possède les propriétés suivantes :
- « Cépages », « Appellation », et « Millésime » de type *owl :DatatypeProperty.*
- « ProduitDans » et « ProduitPar » de type *owl :Objectproperty.*

Les valeurs des propriétés « ProduitDans » et « Cepages » sont communes à l'ensemble des instances de « Muscadet ». La classe « Muscadet » possède une instance, « MaBouteilleDeMuscadet ».

La fonction $\mathcal{V}$ appliquée sur « MaBouteilleDeMuscadet » retournera l'ensemble {*"1998", "melon de bourgogne", "muscadet Sèvre-et-Maine", "muscadet lie"*}.

La fonction $\mathcal{O}$ appliquée à la propriété « Appellation » de l'instance « MaBouteilleDeMuscadet » retournera {*"muscadet Sèvre-et-Maine", "muscadet lie"*}. Cette même fonction appliquée à la propriété « ProduitPar » de « MaBouteilleDeMuscadet » retournera l'instance « DomaineDeTruc ». La fonction $\mathcal{V}$ appliquée à cette instance donnera l'ensemble {*"Domaine de Truc", "LA HAYE FOUASSIERE"*}.

A partir des fonctions d'annotation et d'extraction d'information, nous pouvons maintenant définir les ensembles de termes et données associés aux classes et entités. Nous présenterons les définitions des fonctions σ^0 à partir desquelles nous formerons les relations d'association finales σ'.

Traitement de la hiérarchie des classes

Pour une classe OWL (*owl :Class*) $x \in C$, sa représentation $\sigma_c^0(x)$ est définie ainsi :

$$\sigma_c^0(x) = \mathcal{A}_{id}(x) \cup Termes(\mathcal{A}(x)) \cup \bigcup_{i \in \sigma_c(x)} \left\{ \mathcal{A}_{id}(i) \cup Termes(\mathcal{A}(i) \cup \mathcal{V}(i)) \right\}$$

La représentation d'une classe est constituée de données issues de sa description intensionnelle et aussi de son extension. A partir de la description intensionnelle de la classe, nous prenons en compte son identifiant et les termes extraits à partir de ses annotations (labels, commentaires etc.). Du coté extensionnel, nous sélectionnons pour chacune des instances de la classe, son identifiant, les termes extraits à partir de ses annotations ainsi que les valeurs des propriétés de type *owl :DatatypeProperty.*

Exemple. Pour la classe « Muscadet », les différents constituants de sa représentation sont définis par :
- $\mathcal{A}_{id}(Muscadet) = \{"muscadet"\}$
- $Termes(\mathcal{A}(Muscadet))$ = { *"muscadet", "type vin", "vin blanc", "vin sec",*
  "vin fançais", "vallée loire", "sud nantes", "vin muscadet", "cépage unique", "melon bourgogne", "gamay bourgogne", "melon musqué", "dénomination lie", "seul hiver", "lie fermentation", "mise bouteille"}

– $\bigcup_{i \in \sigma_c(Muscadet)} \{\mathcal{A}_{id}(i) \cup Termes(\mathcal{A}(i) \cup \mathcal{V}(i))\} = \{$ *"MabouteilleDeMuscadet", "1998", "melon bourgogne", "muscadet Sèvre-et-Maine", "muscadet lies"* $\}$

Finalement l'ensemble des termes et données associés à Muscadet sera $\sigma_c^0(Muscadet) = \{$*"muscadet", "type vin", "vin blanc", "vin sec", "vin fançais", "vallée loire", "sud nantes", "vin muscadet", "cépage unique", "melon bourgogne", "gamay bourgogne", "melon musqué", "dénomination lie", "seul hiver", "lie fermentation", "mise bouteille", "MabouteilleDeMuscadet", "1998", "muscadet Sèvre-et-Maine", "muscadet lies"* $\}$.

Traitement de la hiérarchie des propriétés

En OWL, il existe deux types de propriétés : *owl :DatatypeProperty* et *owl :ObjectProperty*. Ces types de propriétés se différencient par leur type d'image. L'image (appelée également objet) d'une propriété *owl :DatatypeProperty* est une valeur de type simple (chaîne de caractères, date, entier, etc.) tandis que l'image d'une propriété *owl :ObjectProperty* est une instance. Afin de prendre en compte ces spécificités, nous définissons une représentation σ_p^0 spécifique à chaque type de propriété.

Pour une propriété de type *owl :DatatypeProperty* $y \in P$, sa représentation $\sigma_p^0(y)$ est définie ainsi :

$$\sigma_p^0(y) = \mathcal{A}_{id}(y) \cup Termes(\mathcal{A}(y)) \cup \bigcup_{pi \in \sigma_p(y)} \{\mathcal{O}(pi)\}$$

Pour une propriété de type *owl :ObjectProperty* $y \in P$, sa représentation $\sigma_p^0(y)$ est définie ainsi :

$$\sigma_p^0(y) = \mathcal{A}_{id}(y) \cup Termes(\mathcal{A}(y)) \cup \bigcup_{pi \in \sigma_p(y)} \{Termes(\mathcal{A}(\mathcal{O}(pi))) \cup \mathcal{V}(\mathcal{O}(pi))\}$$

Pour chacun des deux types de propriété, leur représentation est constituée, comme pour les classes, de l'identifiant (rdf :ID) et des termes extraits des annotations. Pour les données extensionnelles, la représentation des propriétés de type *owl :DatatypeProperty* contient les valeurs prises par ses instances. La représentation d'une propriété de type *owl :ObjectProperty* contient pour chacune de ses instances les valeurs des propriétés de type *owl :DatatypeProperty* prises par l'objet de l'instance ainsi que les termes contenus dans ses annotations.

Exemple. Prenons l'exemple ci-dessous qui précise la définition des propriétés « ProduitPar » et « Cépages ». La première est de type *ObjectProperty* et la seconde de type *DatatypeProperty*.

```
<owl:ObjectProperty rdf:ID="ProduitPar">
  <rdfs:range rdf:resource="#Producteur"/>
  <rdfs:domain rdf:resource="#Vin"/>
  <rdfs:label>Le producteur d'un vin</rdfs:label>
</owl:ObjectProperty>

<owl:DatatypeProperty rdf:about="#Cepages">
```

```
  <rdfs:domain rdf:resource="#Vin"/>
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:DatatypeProperty>
```

A partir de ces définitions de propriétés et des instances contenues dans l'exemple du muscadet présenté précédemment, on peut définir les ensembles de termes et données qui leur sont respectivement associés. Les différents constituants de l'ensemble associé à « ProduitPar » sont définis par :

- $\mathcal{A}_{id}(ProduitPar) = \{"ProduitPar"\}$
- $Termes(\mathcal{A}(ProduitPar)) = \{"producteur\ vin"\}$
- $\bigcup_{pi \in \sigma_p(ProduitPar)} \{Termes(\mathcal{V}(\mathcal{O}(pi)))\} = \{"Domaine\ de\ Truc",\ "LA\ HAYE\ FOUASSIERE"\}$

Les différents constituants de l'ensemble associé à « Cepages » sont définis par :

- $\mathcal{A}_{id}(Cepages) = \{"Cepages"\}$
- $Termes(\mathcal{A}(Cepages)) = \emptyset$
- $\bigcup_{pi \in \sigma_p(y)} \{\mathcal{O}(Cepages)\} = \{"Melon\ de\ Bourgogne"\}$

L'ensemble de termes et données associés à « ProduitPar » sera $\sigma_p^0(ProduitPar) = \{"ProduitPar",\ "producteur\ vin",\ "Domaine\ de\ Truc",\ "LA\ HAYE\ FOUASSIERE"\}$. L'ensemble associé à « Cepages » sera $\sigma_p^0(Cepages) = \{"Cepages",\ "Melon\ de\ Bourgogne"\}$.

Traitements communs

Finalement, afin d'assurer la propriété d'isomorphisme entre les ensembles ordonnés $(C, \leq)$ et $(2^{T_c}, \subseteq)$ (ou $(P, \leq)$ et $(2^{T_p}, \subseteq)$), nous définissons les relations d'association σ' (σ'_c et σ'_p) à partir des ensembles σ^0 (σ_c^0 et σ_p^0) de la manière suivante :

$$\sigma'(x) = \sigma^0(x) \cup \left\{ \bigcup_{x' < x} \sigma^0(x') \right\}$$

4.2.3 Avantages et limites de ces représentations

Notre modèle de représentation des ontologies décrites en OWL présente quelques limites, notamment de par notre modèle de hiérarchie utilisé. En effet, une ontologie OWL n'est pas limitée à des structures hiérarchiques, et ainsi le langage OWL possède une expressivité plus élevée que notre modèle. Cependant, nous ne prenons pas en compte les notions telles que la transitivité, la cardinalité sur les propriétés. Ainsi, l'algorithme d'alignement utilisé par la suite ne pourra pas s'appuyer sur ces types d'information afin d'aligner des ontologies.

Néanmoins, notre modèle a l'avantage d'être simple et plus universel que les modèles complètement dédiés à ce type de représentation. Il permet, par exemple, de prendre en compte n'importe quel type de hiérarchie textuelle. Sa simplicité lui permet également d'avoir une meilleure intelligibilité auprès de l'utilisateur, qui pourra plus facilement comprendre, par la suite, pourquoi deux entités ont été alignées.

4.3 Extraction de l'alignement et post-traitements

A partir de deux hiérarchies préalablement réindexées sur une extension en partie partagée, nous proposons une méthode de découverte d'alignement extensionnelle et asymétrique. En effet, cette méthode, basée sur le modèle de règle d'association, permet d'extraire un alignement implicatif orienté d'une hiérarchie source vers une hiérarchie cible. Les algorithmes de découverte et sélection des règles entre hiérarchies utilisent deux critères de sélection : le premier permet d'assurer l'admissibilité des règles, et le second permet de réduire la redondance dans l'alignement extrait.

Pour obtenir un alignement symétrique et consistant, nous effectuons des étapes de post-traitement. A partir des deux alignements implicatifs (source vers cible et cible vers source), nous fusionnons ces deux alignements et identifions les équivalences. Ensuite, nous proposons une étape de détection et d'élimination des inconsistances. Afin de répondre à certains critères de cardinalité, nous présentons également des filtres permettant de réduire la cardinalité de l'alignement.

Finalement, nous proposons une dernière méthode d'alignement visant à identifier d'éventuels éléments de redondance, non découverts par la méthode extensionnelle. Cette dernière méthode utilise une combinaison de similarités syntaxiques pour la comparaison des entités. Cette méthode s'appuie sur l'alignement préalablement extrait afin de réduire l'espace de recherche et d'éviter de créer des inconsistances.

4.3.1 Découverte de règles entre hiérarchies

Le processus de découverte de règles prend en entrée deux hiérarchies contextualisées. Ces deux hiérarchies ont subi une étape de pré-traitement permettant de les redéfinir sur des extensions partageant une intersection conséquente. Cette étape peut être réalisée par une des méthodes présentées précédemment. La découverte des règles d'association entre entités se fera, comme précisé section 2.3, sur l'ensemble des objets partagés par les deux hiérarchies.

Ce processus étant asymétrique, nous présentons, dans cette section, la découverte de règles d'une hiérarchie source vers une hiérarchie cible. L'alignement final est obtenu par une deuxième exécution en permutant l'ordre des hiérarchies en entrée.

Lors de la recherche de l'alignement implicatif, nous combinons deux critères pour la sélection des règles. Le premier critère permet de vérifier la qualité implicative d'une règle. Le second est un critère structurel assurant une réduction des redondances dans l'ensemble des règles extraites.

Tout d'abord, nous décrirons les données d'entrée avant de présenter les deux critères de sélection des règles. A partir de ces critères, nous étudierons quelles sont les mesures d'intérêt adaptées aux critères et aux données. Finalement, nous exposerons les algorithmes de découverte des règles entre les deux hiérarchies.

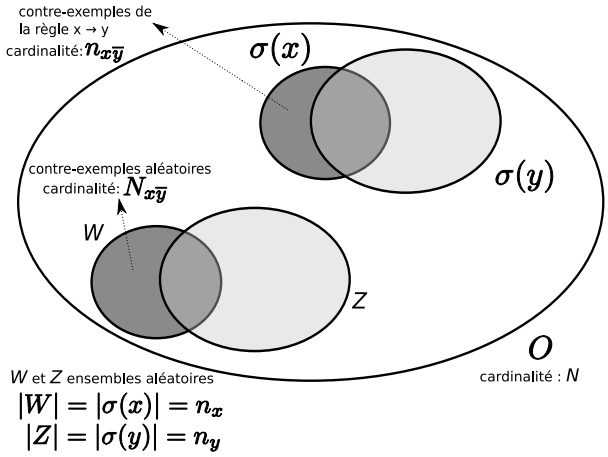


FIG. 4.10 – Intensité d'implication d'une règle $x \rightarrow y$

Données d'entrée

Soit les deux hiérarchies d'entrée suivantes :

$$\mathcal{H}_1 = \{C_1, \leq, \mathcal{A}_1, O_1, \sigma_1)$$
$$\mathcal{H}_2 = \{C_2, \leq, \mathcal{A}_1, O_2, \sigma_2)$$

L'extraction des règles d'association issues d'une entité de C_1 vers une entité de C_2 est réalisée à partir de l'ensemble des objets partagés par les deux hiérarchies, noté $O = O_1 \cap O_2$. Les relations d'association σ_1 et σ_2 sont restreintes à l'ensemble O par la nouvelle relation σ définie par :

$$\sigma(x) = \begin{cases} \sigma_1(x) \cap O_2 \ \ si\ x \in C_1 \\ \sigma_2(x) \cap O_1 \ \ si\ x \in C_2 \end{cases}$$

La notation $x \rightarrow y$ désigne une règle d'association entre une entité $x \in C_1$ vers une entité $y \in C_2$.

Critère d'admissibilité d'une règle

Le premier critère que doit satisfaire une règle $x \rightarrow y$ afin d'être sélectionnée concerne la qualité de la tendance implicative de l'ensemble $\sigma(x)$ vers l'ensemble $\sigma(y)$. Afin de vérifier une telle tendance, de nombreuses mesures sont disponibles (voir section 1.2).

Nous utilisons pour l'évaluation d'une règle $x \rightarrow y$, la mesure d'intensité d'implication dont la modélisation est schématisée sur la figure 4.10 et la valeur est définie par :

$$\varphi(x \rightarrow y) = 1 - Pr\left(N_{x\overline{y}} \leq n_{x\overline{y}}\right)$$

– $n_{x\overline{y}} = |\sigma(x) - \sigma(y)|$ est le nombre d'objets associés à l'entité x mais pas de l'entité y (contre-exemples).

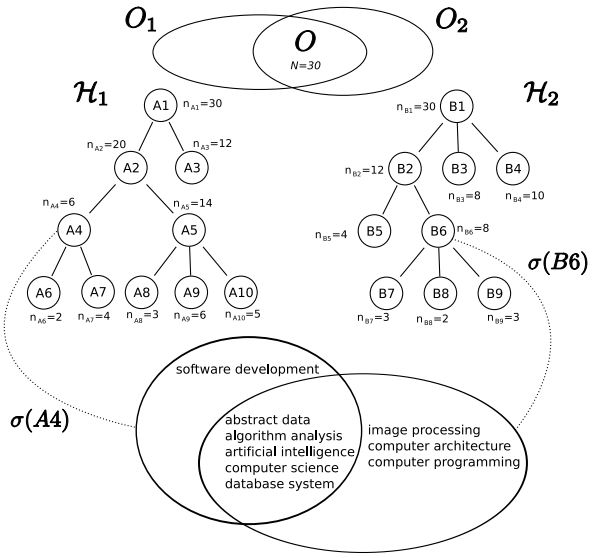


FIG. 4.11 – Evaluation des règles

– $N_{x\overline{y}}$ est le nombre attendu (sous hypothèse d'indépendance) d'objets associés à x mais pas à y.

Sous hypothèse d'indépendance, la probabilité de tirer au hasard un élément de O qui soit associé à x mais pas à y est $P(x\overline{y}) = \frac{|\sigma(x)|\times|O-\sigma(y)|}{|O|^2}$. La variable aléatoire $N_{x\overline{y}}$ a ainsi une espérance $\lambda = \frac{|\sigma(x)|\times|O-\sigma(y)|}{|O|}$.

A partir de sa valeur d'Intensité d'Implication, nous définissons l'admissibilité d'une règle de la manière suivante :

Définition 4.3 *Une règle $x \rightarrow y$, entre les entités $x \in C_1$ et $y \in C_2$, sera* ***admissible*** *au seuil φ_r (niveau de confiance $1 - \varphi_r$) si :*

$$\varphi(x \rightarrow y) \geq \varphi_r \tag{4.1}$$

Exemple. La figure 4.11 présente deux hiérarchies peuplées sur un ensemble commun d'objets, noté O, contenant $N = 30$ éléments. Intéressons-nous à la règle $A4 \rightarrow B6$. Cette règle possède $n_{A4.\overline{B6}} = 1$ contre-exemple (« Software development »), et $n_{A4.B6} = 5$ exemples (« abstract data », « algorithm analysis », « artificial intelligence », « computer science » et « database system »). L'entité « A4 » est associée à $n_{A4} = 6$ objets et l'entité « B6 », à $n_{B6} = 8$ objets. La valeur d'intensité d'implication est donnée par la p-valeur de la loi de Poisson de paramètre $\lambda = \frac{n_{A4}\times(N-n_{B6})}{N} = 4,4$:

$$\varphi(A4 \rightarrow B6) = 1 - e^{-\lambda} \sum_{k=0}^{n_{A4.\overline{B6}}} \frac{\lambda^k}{k!} = 0,97$$

Cette règle sera ainsi admissible si le seuil φ_r utilisé est inférieur ou égal à $0,97$.

Critère de réduction de la redondance

Parmi l'ensemble des règles admissibles entre deux hiérarchies, de nombreuses règles peuvent être redondantes. Cette notion de redondance dans un alignement a été présentée dans la section 2.2.3. Une règle est dite *redondante*, si elle peut être déduite à partir d'une autre règle et éventuellement des connaissances apportées par la relation d'ordre entre entités.

En nous appuyant sur les relations d'ordre respectives des hiérarchies $\mathcal{H}_1$ et $\mathcal{H}_2$, nous proposons un critère de sélection permettant de limiter la redondance lors de l'extraction des règles entre entités.

Définition 4.4 *Une règle $x \rightarrow y$ admissible (c.-à-d. une règle qui respecte la définition 4.3) sera dite* ***significative*** *si elle n'a pas de règle génératrice $x' \rightarrow y'$ ayant une meilleure valeur d'intensité d'implication.*

$$\forall x' \geq x,\ \forall y' \leq y,\ \varphi(x' \rightarrow y') \leq \varphi(x \rightarrow y) \tag{4.2}$$

Exemple. Sur la figure 4.11, les règles potentielles $A4 \rightarrow B7$, $A4 \rightarrow B8$, $A4 \rightarrow B9$, $A2 \rightarrow B8$, $A2 \rightarrow B7$, $A2 \rightarrow B9$, $A1 \rightarrow B6$, $A1 \rightarrow B7$, $A1 \rightarrow B8$ et $A1 \rightarrow B9$ sont plus génératives que la règle admissible $A4 \rightarrow B6$ car elles ont une prémisse plus générale ou une conclusion plus spécifique. La règle $A4 \rightarrow B6$ sera significative, et donc sélectionnée, si aucune de ses règles génératrices n'obtient une valeur d'Intensité d'Implication supérieure.

Ce critère (défintion 4.4) permet de réduire la redondance mais ne la supprime pas totalement. Une règle redondante ayant une meilleure valeur de qualité (c.-à-d. d'Intensité d'Implication) que ses génératrices sera tout de même conservée. Ce choix est justifié par la nature approximative des règles d'association. En effet, il arrive, dans certains cas, qu'une règle ayant une prémisse trop générale ou une conclusion trop spécifique soit sélectionnée.

La figure 4.12 présente deux extraits de hiérarchies : la première classe des appellations de vins selon leur origine géographique, tandis que la deuxième les classe selon leurs caractéristiques organoleptiques. L'extraction d'alignement extensionnel a sélectionné les règles *Muscadet* $\rightarrow$ *vin blanc sec* et *vin du pays Nantais* $\rightarrow$ *vin blanc sec* valuées respectivement à $0,99$ et $0,97$ par l'Intensité d'Implication. La dernière règle, possède une prémisse plus générale que la première. Cependant cette dernière règle, même si elle est significative, n'est pas sémantiquement correcte car les vins du pays Nantais ne sont pas tous des vins blancs secs. En effet, il existe des vins du pays Nantais, appelés vins du pays d'Ancenis, qui peuvent être rouges ou rosés. Cette règle a été sélectionnée car la part des termes apportée par l'entité « vin de pays d'Ancenis » dans l'entité plus générale « vin du pays Nantais » est marginale par rapport à celle des termes de l'entité « Muscadet ».

Mesures d'intérêt appropriées

Afin de quantifier la qualité de la tendance implicative entre les ensembles d'objets associés à deux entités, nous avons choisi d'utiliser la mesure d'Intensité d'Implication. Il existe cependant de nombreuses mesures d'intérêt disponibles

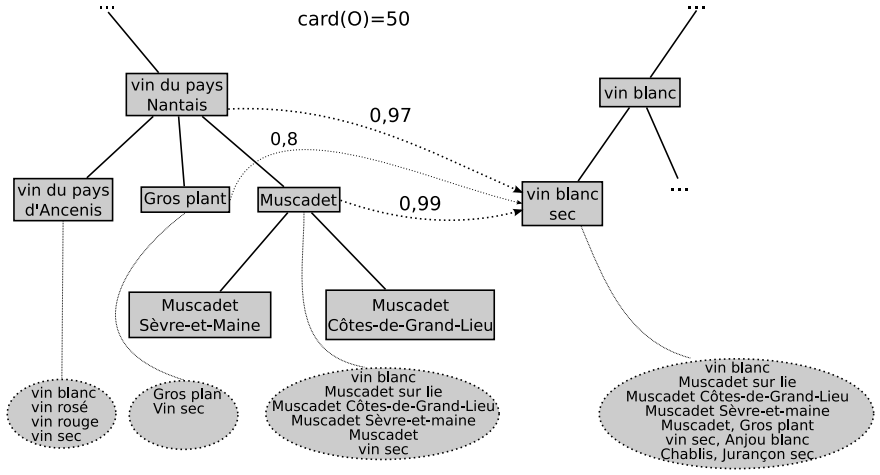


FIG. 4.12 – Exemple de règle trop spécialisée

dans la littérature permettant d'examiner différents aspects des règles. Ces mesures possèdent, selon l'aspect étudié, des propriétés spécifiques et donc des comportement différents. Dans le contexte d'AROMA, certaines familles de mesures sont adaptées, d'autres ne le sont pas. En nous appuyant sur la classification de J. Blanchard [Bla05], nous étudions quelles sont les mesures appropriées à notre méthode.

La classification des mesures d'intérêt des règles d'association montre que, selon leur portée, certaines mesures sont plus adaptées pour évaluer la tendance implicative entre deux ensembles. Ces mesures doivent être asymétriques dans le sens qu'étant données une règle $x \rightarrow y$ et sa réciproque $y \rightarrow x$, la mesure doit évaluer la règle $x \rightarrow y$ avec une meilleure valeur que sa réciproque $y \rightarrow x$ si $n_x < n_y$. Les mesures de quasi-conjonction (les similarités) ou de quasi-équivalence ne sont, dans ce cas, pas adaptées étant donné leur caractère symétrique. Nous rentiendrons ainsi les mesures dont la portée est seulement la règle et les mesures de quasi-implication.

Afin d'optimiser l'efficacité du critère de réduction de redondance 4.2, la mesure d'intérêt utilisée doit, entre autre, favoriser les règles ayant une conclusion la plus spécifique possible étant donnée une prémisse. Considérons les deux règles $x \rightarrow y$ et $x \rightarrow y'$ avec $y' < y$ (illustrées figure 4.13), les cardinalités n_x, n_y, $n_{y'}$, n_{xy}, $n_{xy'}$, n ainsi que les contraintes suivantes :

- $n_{y'} < n_y$ $(\sigma(y') \subset \sigma(y))$.
- $n_{xy'} \leq n_{xy}$ $(\sigma(x) \cap \sigma(y') \subseteq \sigma(x) \cap \sigma(y))$.

La majorité des mesures d'écart à l'équilibre ne dépend que de n_x et de n_{xy} (respectivement $n_{xy'}$) et évalue l'écart entre n_{xy} (resp. $n_{xy'}$) et $n_x/2$. Comme $n_{xy'} \leq n_{xy}$, une telle mesure, notée m_e, va privilégier la règle $x \rightarrow y$ par rapport à $x \rightarrow y'$ et ainsi $m_e(x \rightarrow y) \geq m_e(x \rightarrow y')$. Dans le cas d'une spécialisation parfaite (c.-à-d. si $n_{xy} = n_{xy'}$), les deux règles obtiendront la même valeur de qualité $(m_e(x \rightarrow y) = m_e(x \rightarrow y'))$.

Les mesures d'écart à l'indépendance dépendent en plus des cardinalités

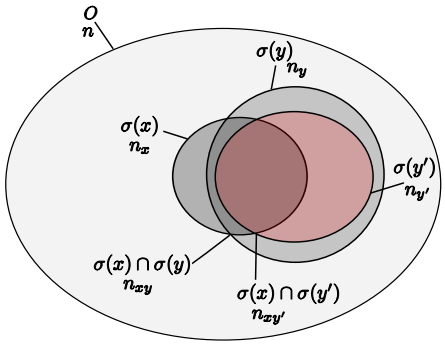


FIG. 4.13 – Spécialisation de la conclusion

n_y (resp. $n_{y'}$) et n et évaluent l'écart entre n_{xy} (resp. $n_{xy'}$) et $n_x.n_y/n$ (resp. $n_x.n_{y'}/n$). Comme $n_x.n_y/n < n_x.n_{y'}/n$, une mesure d'écart à l'indépendance m_i privilégiera $x \rightarrow y'$ à $x \rightarrow y : m_i(x \rightarrow y) < m_i(x \rightarrow y')$ dans le cas d'une spécialisation parfaite.

Les mesures asymétriques d'écart à l'indépendance (dont la portée est la quasi-implication ou la règle) sont, de par leur propriétés, plus adaptées à l'évaluation des règles non-redondantes. En effet ces mesures permettent d'une part de quantifier la tendance implicative (entre les ensembles d'objets associés aux deux entités) et d'autre part, de favoriser les règles ayant une conclusion spécifique parce qu'elles sont décroissantes lorsque la taille de la conclusion augmente et que les autres paramètres restent identiques.

Algorithmes de découverte de règles

AROMA effectue une recherche descendante des règles d'association en exploitant la relation d'ordre partiel, permettant ainsi de réduire le temps de calcul. Elle utilise deux algorithmes pour découvir les règles entre les entités d'une hiérarchie source $\mathcal{H}_1$ et une hiérarchie cible $\mathcal{H}_2$. Le premier algorithme dirige la recherche globale et effectue la descente dans la hiérarchie source. Le deuxième permet de trouver les meilleures règles satisfaisant les critères de sélection pour une prémisse donnée.

Le premier algorithme (algorithme 1) prend en entrée une entité x de la hiérarchie source ($\mathcal{H}_1$) et un ensemble d'entités cibles, *ConclusionSet* issus de $\mathcal{H}_2$. Pour chaque entité $y \in ConclusionSet$, l'intersection entre les ensembles de termes ou données de x et y est évaluée. Si cette intersection est l'ensemble vide (c.-à-d. $support(xy) = |\sigma(x) \cap \sigma(y)| = 0$) alors l'ensemble des descendants de y (c.-à-d. le sous-arbre contenant y comme racine) ne sera pas considéré. Dans l'autre cas, la recherche des règles valides entre x et les entités du sous-arbre de y est effectuée. Cette procédure est lancée de manière récursive sur tous les descendants de x.

Pour une entité source x et une entité cible y, le deuxième algorithme (algorithme 2) permet de trouver toutes les règles de la forme $x \rightarrow y'$ (avec $y' \leq y$) satisfaisant les deux critères de sélection (équations (4.1) et (4.2)). L'algorithme

Algorithme 1 fonction spécialisePrémisse

Entrées : $x \in C_1$,
$ConclusionSet \subset C_2$,
$ImplicationSet$ (ensemble des règles sélectionnées)
Sorties : L'ensemble des règles d'association entre le sous-arbre ayant comme racine x et chacun des sous-arbres ayant une racine dans $ConclusionSet$

```
pour tout y ∈ ConclusionSet faire
  si σ(x) ∩ σ(y) ≠ ∅ alors
    spécialiseConclusion(x, y, ImplicationSet, φ_r)
  sinon
    ConclusionSet := ConclusionSet − {y}
  fin si
fin pour
pour tout fils ∈ lesFils(x) faire
  spécialisePrémisse(fils, CopieDe(ConclusionSet), ImplicationSet)
fin pour
retourner ImplicationSet
```

arrête la recherche de règles dans un sous-arbre si la valeur de l'intensité d'implication devient trop petite : si la valeur $\varphi(x \rightarrow x \cap y)$ est inférieure au seuil de sélection des règles φ_r, alors aucune règle ayant une spécialisation de y ne pourra avoir une valeur supérieure à ce seuil. Cette propriété est vraie pour l'ensemble des mesures asymétriques d'écart à l'équilibre et d'écart à l'indépendance. Les racines des hiérarchies ne sont pas évaluées car tous les termes ou données sont associés à ces entités. La valeur d'intensité d'implication de telles règles (ayant une prémisse ou conclusion racine) est soit non-définie, soit égale à 0.

4.3.2 Post-traitements

A partir des deux alignements implicatifs $A_{\mathcal{H}_1 \rightarrow \mathcal{H}_2}$ et $A_{\mathcal{H}_2 \rightarrow \mathcal{H}_1}$ calculés, nous déduisons un alignement symétrique A (c.-à-d. contenant des éléments de correspondance de type implication ($\Rightarrow$ et $\Leftarrow$) et de type équivalence). Cet alignement symétrique est réalisé par la fusion des deux alignements implicatifs et la déduction des équivalences. A la fin de cette première étape, l'alignement pouvant contenir des inconsistances, nous proposons un filtre permettant leur détection et leur élimination. La cardinalité de l'alignement final obtenu est $0, n-0, n$, c.-à-d. qu'une entité issue de $\mathcal{H}_1$ peut être associée à 0, 1 ou plusieurs entités issues de la hiérarchie $\mathcal{H}_2$ et vice-versa. Afin de pouvoir produire des alignements de cardinalité restreinte, nous proposons également une démarche de réduction de la cardinalité.

Déduction des équivalences

La déduction des entités en relation d'équivalence est réalisée de manière directe, c.-à.-d. que nous n'utilisons pas les propriétés de déduction exposées dans la section 2.2.2. A partir des ensembles $V_{\mathcal{H}_1 \rightarrow \mathcal{H}_2}$ et $V_{\mathcal{H}_2 \rightarrow \mathcal{H}_1}$, l'alignement symétrique fusionné V (c.-à-d. contenant les relations d'équivalence déduites et

Algorithme 2 fonction spécialiseConclusion

Entrées : $x \in C_1$,
$y \in C_2$,
$ImplicationSet$ (ensemble des règles sélectionnées) et
φ_r (seuil de sélection des règles)
Sorties : La valeur de l'intensité d'implication de la meilleure règle ayant x comme prémisse et y' (avec $y' \leq y$) comme conclusion

```
meilleurFils := faux
φ_courante := φ(x → y)
φ_max := φ_courante
si φ(x → x ∩ y) ≥ φ_r alors
  pour tout fils ∈ lesFils(y) faire
    φ_fils ← spécialiseConclusion(x, y)
    si φ_courante ≤ φ_fils alors
      meilleurFils := vrai
      φ_max := max(φ_max, φ_fils)
    fin si
  fin pour
  si ¬meilleurFils ∧ (φ_courante ≥ φ_r) ∧
  (∀x' → y' ∈ ImplicationSet, x ≤ x', y' ≤ y, φ(x → y) ≤ φ(x' → y')) alors
    ImplicationSet := ImplicationSet ∪ {x → y}
  fin si
fin si
retourner φ_max
```

également les relations d'implication) est défini par :

$$\begin{aligned} V = \{x \Leftrightarrow y | x \Rightarrow y \in V_{\mathcal{H}_1 \to \mathcal{H}_2} \wedge x \Leftarrow y \in V_{\mathcal{H}_2 \to \mathcal{H}_1}\} \cup \\ \{x \Rightarrow y | x \Rightarrow y \in V_{\mathcal{H}_1 \to \mathcal{H}_2} \wedge x \Leftarrow y \notin V_{\mathcal{H}_2 \to \mathcal{H}_1}\} \cup \\ \{x \Leftarrow y | x \Rightarrow y \notin V_{\mathcal{H}_1 \to \mathcal{H}_2} \wedge x \Leftarrow y \in V_{\mathcal{H}_2 \to \mathcal{H}_1}\} \end{aligned} \quad (4.3)$$

La mesure de qualité évaluant la pertinence d'une équivalence est déduite à partir des valeurs d'intensité d'implication des deux implications dont elle est issue. Nous utilisons, pour cela, la moyenne géométrique des valeurs d'Intensité d'Implication :

$$\varphi(x \leftrightarrow y) = \sqrt{\varphi(x \to y) \times \varphi(y \to x)}$$

La fonction de qualité q associée à l'alignement est alors définie par :

$$\begin{aligned} q = \{(x \Rightarrow y, \varphi(x \to y)) | x \Rightarrow y \in V\} \cup \\ \{(x \Leftarrow y, \varphi(y \to x)) | x \Leftarrow y \in V\} \cup \\ \{(x \Leftrightarrow y, \varphi(x \leftrightarrow y)) | x \Leftrightarrow y \in V\} \end{aligned} \quad (4.4)$$

Elimination des inconsistances

Le processus d'élimination des inconsistances permet d'obtenir un alignement consistant. Tout d'abord, nous proposons de détecter les couples

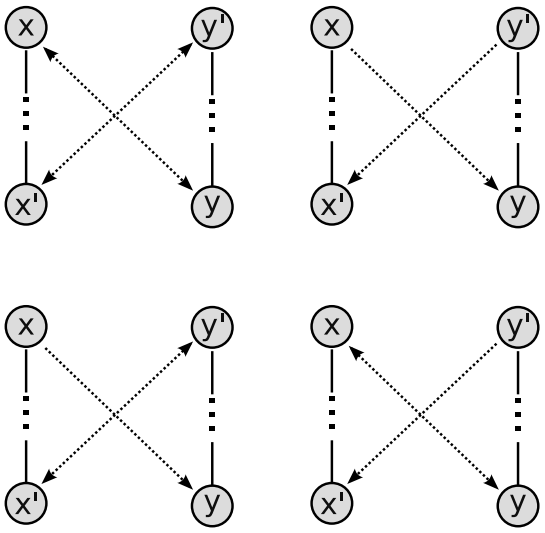


FIG. 4.14 – les quatre schémas d'inconsistance possibles

d'éléments de correspondance en contradiction dans un alignement symétrique (c.-à-d. composé d'éléments de correspondance de type relation d'implication $\Rightarrow$, $\Leftarrow$ et relation d'équivalence $\Leftrightarrow$). Ensuite, en fonction du type d'inconsistance détectée, nous établissons des règles visant à éliminer ces contradictions soit en supprimant l'un des deux éléments de correspondance en cause, soit par construction d'un nouvel élement de correspondance remplaçant les deux éléments en contradiction.

A partir de la définition d'une contradiction (déf. 2.13), nous avons recensé quatre schémas possibles de contradiction dans un alignement. Ces schémas sont illustrés sur la figure 4.14. Dans les quatre cas, nous avons quatre entités respectant les contraintes suivantes : $x' \leq x$, $y \leq y'$ et $x \neq x' \vee y \neq y'$. La dernière contrainte permet de spécialiser la situation à trois concepts : soit $x = x'$, soit $y = y'$. Dans chaque cas, nous avons deux éléments de correspondance, respectivement $x\mathcal{R}y$ et $x'\mathcal{R}'y'$.

Dans le premier cas, les deux éléments de correspondance en contradiction sont de type équivalence : $x \Leftrightarrow y$ et $x' \Leftrightarrow y'$. Dans ce cas, nous retiendrons l'élément qui a la meilleure valeur de qualité. Si les deux éléments de correspondance ont la même valeur de qualité alors nous les généraliserons par la relation $x \Leftrightarrow y'$.

Dans le deuxième cas, les deux éléments de correspondance en contradiction sont des implications : $x \Rightarrow y$ et $x' \Leftarrow y'$. Dans ce cas, nous remplaçons les deux relations en contradiction, $x \Rightarrow y$ et $x' \Leftarrow y'$, par l'équivalence $x \Leftrightarrow y'$.

Dans les deux derniers cas, une relation d'implication est en contradiction avec une relation de type équivalence. Dans ces deux cas, nous choisissons de garder l'élément de correspondance de type équivalence et de supprimer l'implication.

Réduction des cardinalités et restriction asymétrique

Les alignements calculés par AROMA sont symétriques et de cardinalité $0,n-0,n$, c.-à-d. qu'une entité de la hiérarchie source (resp. cible) peut être associée à zéro, une, ou plusieurs entités de la hiérarchie cible (resp. source). Cependant, en fonction de l'application visée, on peut préférer des alignements de nature asymétrique, car seules les relations d'une hiérarchie source vers une hiérarchie cible nous intéressent. On peut également vouloir obtenir un alignement fonctionnel, c.-à-d. qu'une entité de la hiérarchie source sera associée à, au plus, une entité entité de la hiérarchie cible.

La réduction d'un alignement A à l'une de ses composantes asymétriques $V_{\Rightarrow}$ (resp. $V_{\Leftarrow}$) est obtenu facilement en éliminant les éléments de correspondance dont la relation est l'implication $\Leftarrow$ (resp. $\Rightarrow$).

$$V_{\Rightarrow} = \{x\mathcal{R}y | x\mathcal{R}y \in V \wedge (\mathcal{R} = \Rightarrow \vee \mathcal{R} = \Leftrightarrow)\}$$

$$V_{\Leftarrow} = \{x\mathcal{R}y | x\mathcal{R}y \in V \wedge (\mathcal{R} = \Leftarrow \vee \mathcal{R} = \Leftrightarrow)\}$$

Afin de rendre un alignement fonctionnel, nous adoptons le principe suivant : pour chaque entité de la hiérarchie source, nous gardons l'élément de correspondance qui a la meilleure valeur de qualité. Cependant, il peut arriver que plusieurs élements de correspondance partagent la meilleure valeur de qualité. Dans ce cas, afin de les départager, nous utilisons une seconde mesure ayant des critères d'évaluation différents de la première. Comme la méthode utilise l'Intensité d'Implication qui est une mesure statistique d'écart à l'indépendance, nous utiliserons la confiance, une mesure d'écart à l'équilibre qui est quant à elle, descriptive. La confiance d'une règle $x \rightarrow y$ est définie par :

$$conf(x \rightarrow y) = \frac{|\sigma'(x) \cap \sigma'(y)|}{|\sigma'(x)|}$$

Dans le cas d'éléments de correspondance $x \Leftrightarrow y$ en relation d'équivalence, nous utiliserons comme pour l'intensité d'implication, la moyenne géométrique des qualités obtenues avec la confiance pour chacune des règles $x \rightarrow y$ et $y \rightarrow x$ qui ont permis d'obtenir l'équivalence. Cette moyenne géométrique des valeurs de confiance donne un indice de similarité connu sous le nom d'indice d'Ochiai [Och57].

$$conf(x \leftrightarrow y) = \sqrt{conf(x \rightarrow y) \times conf(y \rightarrow x)}$$

Exemple. Sur la figure 4.15, l'entité $A8$ est en correspondance avec $B8$ et $B6$. Les deux éléments de correspondance sous-jacents sont tous deux évalués à 1 par l'intensité d'implication. Afin de choisir, un des deux éléments de correspondance, nous calculons la confiance des règles $A8 \rightarrow B8$ et $A8 \rightarrow B6$: $conf(A8 \rightarrow B8) = n_{A8.B8}/n_{A8} = 1$ et $conf(A8 \rightarrow B6) = n_{A8.B6}/n_{A8} = 0,67$. A partir de cette deuxième évaluation, nous pouvons ainsi garder l'élément de correspondance $A8 \rightarrow B8$.

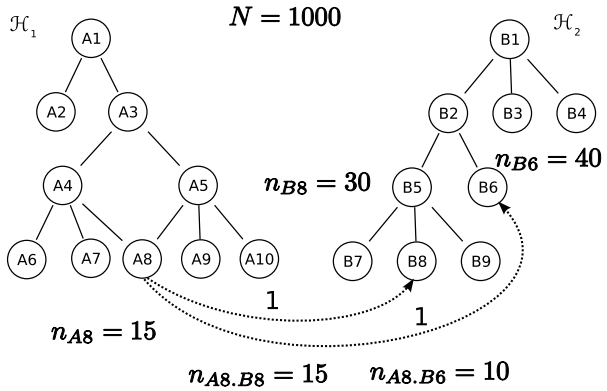


FIG. 4.15 – Choix entre deux éléments de correspondance de qualité identique

4.3.3 Similarité syntaxique sur description intensionnelle

L'extraction des règles d'alignement entre deux entités n'est possible que lorsque ces dernières sont suffisamment décrites dans les corpus associés aux hiérarchies afin que l'éventuelle tendance implicative puisse être validée statistiquement. Ainsi, il arrive assez fréquemment que certaines entités, généralement feuilles d'une hiérarchie, soient associées à aucun ou très peu de termes ou données. Dans ce cas, aucune relation d'alignement n'a pu être évaluée avec ces entités alors qu'elles pourraient avoir une correspondance avec d'autres entités dans la seconde hiérarchie. Afin de pallier ce type de problème et d'améliorer l'alignement produit, nous proposons d'utiliser un second algorithme basé sur une comparaison des noms des entités (et de leurs descriptions). Nous proposons également d'utiliser l'alignement préalablement extrait afin de rendre la recherche rapide et cohérente.

Le principe de l'algorithme proposé est de rechercher pour chaque entité n'apparaissant pas dans l'alignement d'entrée, ses entités généralisantes les plus spécifiques en utilisant la relation d'ordre partiel et l'alignement fourni. Ensuite, cette entité sera comparée aux descendantes des entités généralisantes. La similarité utilisée pour la comparaison est une combinaison de similarités entre chaînes de caractères appliquées aux différents niveaux de description des entités.

Similarité syntaxique entre entités

En fonction des types de hiérarchies considérées (ontologies OWL, annuaires de sites web, catalogues hiérarchiques, etc.), les entités peuvent être décrites de manière plus ou moins expressive. Notamment, au niveau terminologique, une entité possède en général un nom, mais elle peut être également associée à un identifiant, des commentaires, etc. Nous proposons dans AROMA de prendre en compte ces différents niveaux d'information terminologique en combinant les similarités obtenues entre les différentes informations textuelles décrivant les entités au niveau du schéma. Ce principe est utilisé dans l'algorithme ASCO [Bac06].

Dans le cas générique, nous utilisons, pour comparer les noms de deux entités a et b, la similarité de JaroWinkler (définie équation 3.4.1). Cette similarité donne de bons résultats et est particulièrement adaptée pour comparer des chaînes de caractères composées d'un ou quelques mots [CRF03].

$$sim_n(a,b) = JaroWinkler(nom(a), nom(b))$$

Afin de prendre en compte de l'information structurelle, cette similarité peut être hybridée en comparant également la concaténation des noms reliant l'entité racine de la hiérarchie à l'entité considérée. Ce principe est une similarité basée sur le chemins de noms (appelée, name path-based similarity, en anglais).

Dans le cas, où les entités sont également décrites par des commentaires, on prendra en compte une similarité syntaxique basée sur les ensembles de mots. Nous proposons d'utiliser la méthode vectorielle largement utilisée en recherche d'information et qui donne de bons résultats lorsque les descriptions textuelles sont relativement longues (c.-à-d. lorsqu'elles contiennent une à plusieurs phrases). Cette méthode représente chaque commentaire sous forme vectorielle où les dimensions représentent les mots contenus dans les commentaires des entités a et b et les composantes sont les valeurs données par la mesure $TF.IDF$ (présentée section 3.5.1). Soit $\overrightarrow{Vect(t)}$, la représentation vectorielle d'un texte t. La similarité entre deux entités sur la base de leurs commentaires sera donnée par la valeur du cosinus de l'angle formé par leurs vecteurs respectifs :

$$sim_c(a,b) = cos(\overrightarrow{Vect(commentaires(a))}, \overrightarrow{Vect(commentaires(b))})$$

Finalement, dans le cas, où plusieurs de ces similarités sont utilisables, on agrège leurs résultats par une moyenne éventuellement pondérée par la confiance que l'on souhaite donner à chaque description.

Dans le cas générique (hiérarchies textuelles), la similarité syntaxique, donnée par l'équation 4.5, sera la moyenne des similarités entre les noms et les chemins de noms.

$$sim(a,b) = \frac{sim_n(a,b) + sim_{np}(a,b)}{2} \tag{4.5}$$

Dans le cas d'ontologies décrites en RDFS/OWL, la similarité syntaxique prendra également en compte les similarités entre identifiants et commentaires. La similarité entre identifiants sim_{id} est, comme sim_n, basée sur la similarité de JaroWinkler.

$$sim(a,b) = \frac{sim_n(a,b) + sim_{np}(a,b) + sim_{id}(a,b) + sim_c(a,b)}{4}$$

Algorithme d'alignement syntaxique

Etant donnés deux hiérarchies $\mathcal{H}_1$ et $\mathcal{H}_2$, et un alignement $A(V,q)$ (voir défintion 2.7), nous définissons les primitives suivantes :

- $NonAlignes(C_x, V) = \{x |\ \nexists y,\ x\mathcal{R}y \in V\}$: cette fonction retourne l'ensemble des entités de C_x ($x \in \{1,2\}$) qui n'apparaissent pas dans V.

- $fils(X) = \bigcup_{x \in X} \{x' | x' \prec x\}$: cette fonction retourne l'ensemble des descendants directs par $\leq$ de l'ensemble des entités de X.
- $peres(x) = \{x' | x \prec x'\}$: cette fonction retourne l'ensemble des ascendants directs de l'entité x par $\leq$.
- $ppag(x)$: Cette fonction, définie par l'algorithme 4.3.3, recherche récursivement parmi les ascendants de x, les plus petits généralisants qui sont en relation d'implication ou d'équivalence dans V et retourne leurs images.

Algorithme 3 ppag : Découverte des plus petits alignés généralisants

Entrées : une entité x (et sa hiérarchie $\mathcal{H}_x$), un alignement $A = (V, q)$.
Sorties : le ppag de x étant donné V s'il existe, $\emptyset$ sinon.

```
PPAG := ∅
PERES := peres(x)
si PERES = ∅ alors
  retourner ∅
fin si
pour tout x' ∈ PERES faire
  PPAG := PPAG ∪ {y' | (x', y', R) ∈ V ∧ R ∈ {→, ↔}}
fin pour
si PPAG = ∅ alors
  pour tout a' ∈ PERES faire
    PPAG := PPAG ∪ ppag(x')
  fin pour
fin si
retourner PPAG
```

Pour chaque entité, notée x qui n'intervient dans aucune relation d'alignement, l'algorithme sélectionne les images y' des ascendants les plus spécifiques de x, notés x' qui interviennent dans une relation d'alignement de type équivalence ou implication notée, $x' \Leftrightarrow y'$ ou $x' \Rightarrow y'$. Ensuite, les similarités syntaxiques sont évaluées entre l'entité x et chacune des entités descendantes directes de y' qui ne figurent pas dans l'alignement de départ $A(V, q)$. Une relation d'équivalence entre x et $y \leq y'$ sera retenue si la valeur de la similarité $sim(x, y)$ est supérieure à une valeur seuil fixée et si l'entité y est celle, parmi toutes les candidates, qui maximise cette valeur de similarité. Si aucune similarité n'a une valeur supérieure au seuil ($sim(x, y) < seuil_{sim}$) alors la recherche se poursuit sur l'ensemble des descendants directs des y.

Sur la figure 4.16, la mesure de similarité syntaxique sera évaluée d'abord entre l'entité a et les entités $b1$ et $b2$, puis éventuellement entre a et les entités $b3$, $b4$ et $b5$ (qui sont toutes descendantes de b'). L'évaluation de la similarité entre en a et b'' n'est pas envisagée car une relation entre ces deux entités ne serait pas appuyée par la relation $a' \leftrightarrow b'$.

Exemple. La figure 4.17 présente deux hiérarchies organisant les cours proposés par deux universités. Un alignement extensionnel a été préalablement calculé entre ces deux hiérarchies. L'entité « Philosophie » de la première hiérarchie (« Université truc ») n'a pas été alignée. Nous recherchons alors une éventuelle correspondance en utilisant une similarité syntaxique sur les chaînes de caractères. Si la méthode de calcul de similarité syntaxique considère seulement

Algorithme 4 syntaxeAligne : Découverte d'un alignement par approche syntaxique

Entrées : deux hiérarchies $\mathcal{H}_1$ et $\mathcal{H}_2$, un alignement $A = (V, q)$, un seuil de sélection $seuil_{sim}$

Sorties : un alignement $A'(V', q')$ contenant de nouvelles relations d'équivalence.

pour tout $x \in NonAlignes(C_1, V)$ **faire**
  $CANDIDATS := fils(ppag(x, V)) \cap NonAlignes(C_2, V)$
  tantque $CANDIDATS \neq \emptyset$ **faire**
    $meilleur := maxSim(a, CANDIDATS)$
    si $meilleur \neq null \wedge sim(x, meilleur) > seuil_{sim}$ **alors**
      $nouvEquiv := (x, meilleur, \leftrightarrow)$
      $V' := V' \cup \{nouvEquiv\}$
      $q' := q' \cup \{(nouvEqui, sim(a, meilleur))\}$
    sinon
      $CANDIDATS := fils(CANDIDATS) \cap NonAlignes(C_2, V)$
    fin si
  fin tantque
fin pour
$A'' = (V'', q'') := syntaxeAligne(\mathcal{H}_2, \mathcal{H}_1, (V \cup V', q \cup q'), seuil_{sim})$
retourner $(V' \cup V'', q' \cup q'')$

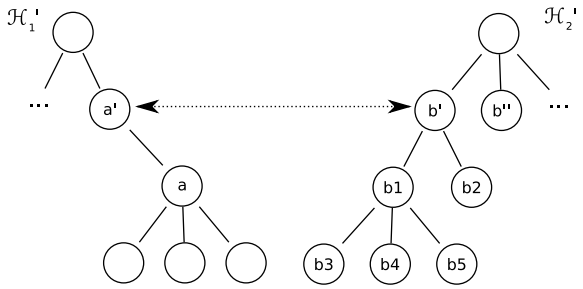


FIG. 4.16 – Exemple d'alignement syntaxique

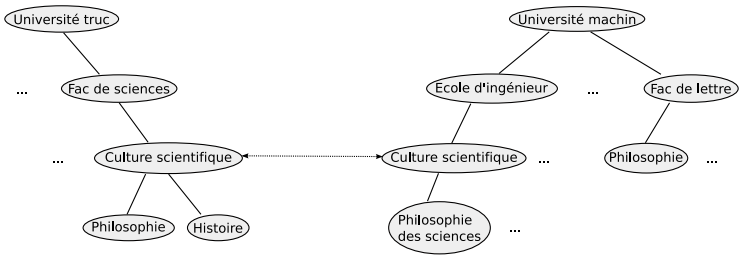


FIG. 4.17 – Extrait d'alignement entre deux catalogues de cours

leur nom et ne prend pas en compte l'alignement préalablement calculé, l'entité « philosophie » de la première hiérarchie sera alignée avec l'entité « philosophie » de la deuxième hiérarchie car leur nom est strictement identique. Cependant, cet élément de correspondance n'est pas sémantiquement valide car l'entité « philosophie » désigne dans la première hiérarchie, un cours de culture scientifique alors que dans la deuxième, l'entité, portant le même nom, désigne une formation de la fac de lettres. Si l'on utilise l'élément de correspondance « culture scientifique » ⇔ « culture scientifique », alors l'entité « philosophie » sera probablement alignée avec l'entité « philosophie des sciences ».

Conclusion

Nous avons présenté dans ce chapitre, notre méthode d'alignement de hiérarchies contextualisées. Cette méthode, appelée AROMA, repose en majeure partie sur la description extensionnelle des hiérarchies. Le processus d'alignement se déroule en trois phases principales : (1) le pré-traitement des hiérarchies qui permet de les redéfinir sur un ensemble de termes communs ; (2) la découverte de règles d'association entre les entités issues des deux hiérarchies ; (3) le post-traitement des résultats qui d'obtenir un alignement consistant et minimal (selon un critère de redondance).

La première phase de pré-traitement est dépendante du type de hiérarchies traitées. Nous avons proposé, dans la première section, deux méthodes de pré-traitement : l'une dédiée aux hiérarchies textuelles et une autre adaptée aux ontologies RDFS/OWL. Tandis que les deux méthodes reposent sur une approche de TAL, la première méthode s'appuie également sur une sélection statistique des termes.

Le processus de découverte de l'alignement repose sur l'extraction et l'évaluation des règles d'association entre entités. Une règle d'association admissible, signifie que le vocabulaire (et les données) associé à une entité source tend à être inclus dans le vocabulaire de l'entité cible. Si une règle admissible est également significative (c.-à-d. qu'elle n'a pas de règle plus générique ayant une meilleure valeur de qualité), nous retiendrons une relation d'implication entre ces deux entités.

L'étape de post-traitement permet tout d'abord de déduire des équivalences à partir de l'alignement implicatif. Ensuite, il est réalisé une détection et une suppression des éléments de correspondance en contradiction. Nous proposons également un filtre de réduction de la cardinalité de l'alignement. Finalement, une dernière méthode basée sur des similarités syntaxiques permet de découvrir d'éventuels éléments de correspondances non détectés par la méthode extensionnelle. Cette dernière méthode a l'originalité de s'appuyer sur l'alignement préalablement produit afin d'éviter de produire des inconsistances.

Notre méthode a l'avantage de tirer profit à la fois des descriptions intensionnelles et extensionnelles d'une hiérarchie afin d'extraire des alignements consistants et minimaux (ne contenant pas de redondance). Elle a en outre l'originalité de détecter des éléments de correspondance de type implication qui sont très rarement pris en compte par les méthodes disponibles.

5 Mesures pour l'évaluation

Sommaire

Introduction

Pour évaluer la performance d'une méthode d'alignement, on utilise un alignement de référence, généralement construit à la main, et on compare les résultats produits par une méthode par rapport à cet alignement de référence. Le modèle de comparaison utilisé est celui largement utilisé en recherche d'information basé notamment sur les mesures de précision et de rappel. Cependant, ce modèle présente des limites quant à l'évaluation des méthodes d'alignement car il ne prend pas en compte la sémantique des structures alignées et des éléments de correspondance. Des modèles plus adaptés ont été proposés dans la littérature [EE05], [Euz07]. Même s'ils permettent de mieux prendre en compte la sémantique de l'alignement, ils supportent mal la présence de redondance et également la présence simultanée d'éléments de correspondance de type équivalence et implication.

Nous proposons dans ce chapitre, de rappeler, dans un premier temps, le principe du modèle classique d'évaluation. Nous étudions ensuite le modèle sémantique introduit dans [Euz07] et analysons ses limites. Finalement, nous présentons une adaptation palliant les limites de ce dernier modèle.

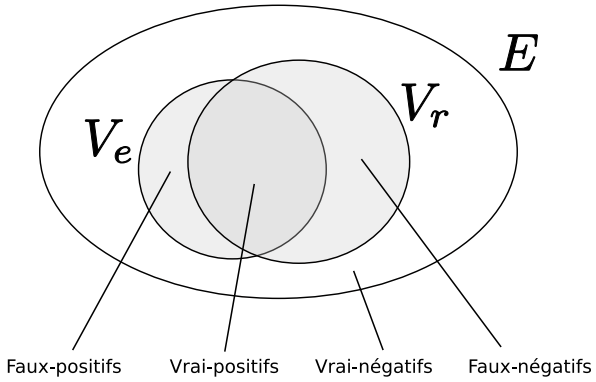


FIG. 5.1 – Diagramme de Venn du modèle d'évaluation classique

	pertinents	non pertinents	
trouvés	$\|V_e \cap V_r\|$ vrai-positifs	$\|V_e - V_r\|$ faux-positifs	$\|V_e\|$
non trouvés	$\|V_r - V_e\|$ faux-négatifs	$\|(E - V_r) - V_e\|$ vrai-négatifs	$\|E - V_e\|$
	$\|V_r\|$	$\|E - V_r\|$	

TAB. 5.1 – Contingence des ensembles V_e et V_r

5.1 Modèle d'évaluation classique

Soit un alignement de référence $A_r = (V_r, q_r)$ ainsi qu'un alignement à évaluer $A_e = (V_e, q_e)$ produit par la méthode. En principe, la fonction de qualité q_r assigne une valeur de 1 à chaque élément de V_r. Les fonctions de qualité ne sont pas comparées et n'entrent donc pas en compte dans l'évaluation.

A l'accoutumée, les résultats produits par les méthodes d'alignement sont comparés à l'alignement de référence en utilisant des mesures d'évaluation typiques de la recherche d'informations. Les deux mesures de base sont la précision et le rappel. Cette approche modélise le problème de manière ensembliste, voir figure 5.1, et considère les ensembles suivants :

- vrai-positifs : les éléments de correspondance trouvés par la méthode qui sont dans l'alignement de référence.
- faux-positifs : les éléments de correspondance trouvés par la méthode qui ne sont pas dans l'alignement de référence.
- faux-négatifs : les éléments de correspondance de l'alignement de référence qui n'ont pas été trouvés par la méthode.

Les cardinalités des ensembles des vrai-positifs, des faux-positifs et des faux-négatifs sont exprimés à partir de V_e et V_r. La table de « contingence » 5.1 donne ces cardinalités. On représente par E, l'ensemble des éléments de correspondances possibles que l'on peut former à partir des deux hiérarchies et des relations considérées.

5.1.1 Mesures classiques d'évaluation

La précision, notée P, représente la proportion de vrai-positifs parmi l'ensemble des éléments de correspondance trouvés par la méthode. Cette mesure permet de qualifier la pertinence de la méthode d'alignement. Plus la valeur de précision se rapproche de 1, moins la méthode produit de bruit.

$$P(V_e, V_r) = \frac{|V_e \cap V_r|}{|V_e|}$$

Le rappel, noté R, représente la proportion de vrai-positifs parmi l'ensemble des éléments de correspondance contenus dans l'alignement de référence. Cette mesure permet de quantifier la couverture de la méthode d'alignement. Plus le rappel se rapproche de 1, moins la méthode est silencieuse.

$$R(V_e, V_r) = \frac{|V_e \cap V_r|}{|V_r|}$$

En pratique, il est assez facile d'obtenir une des deux valeurs (précision ou rappel) à 1. Pour le rappel, il suffit de retourner tous les élements de correspondance qu'il est possible de générer : on obtiendra ainsi un rappel de 1 mais une précision sans doute mauvaise. Pour la précision, il suffit de retourner seulement quelques éléments de correspondance dont on est sûr qu'ils soient corrects. On obtient alors une précision, à priori, de 1 mais un rappel faible. C'est pourquoi ces mesures doivent être toujours utilisées conjointement pour obtenir une bonne idée de la performance de la méthode.

Ces deux mesures de base peuvent être combinées par la F-mesure [vR79]. La F-mesure, notée F, représente la moyenne harmonique entre la précision et le rappel. Elle représente également la mesure de similarité de Dice [Dic45] entre les ensembles V_e et V_r.

$$\begin{aligned} F(V_e, V_r) &= \frac{2 \times P(V_e, V_r) \times R(V_e, V_r)}{P(V_e, V_r) + R(V_e, V_r)} \\ &= \frac{2 \times |V_e \cap V_r|}{|V_e| + |V_r|} \end{aligned}$$

5.1.2 Limites

Ce modèle d'évaluation est convenable lorsque les alignements (A_e et A_r) ne considèrent que la relation d'équivalence $\Leftrightarrow$. Cependant, lorsque l'on prend en compte les relations d'implication $\Rightarrow$ et $\Leftarrow$, des déductions à partir des alignements et des hiérarchies deviennent possibles rendant ce modèle inadéquat. Ces limites proviennent de deux faits :

1. Le modèle ne prend pas en compte le fait qu'une double implication $x \Rightarrow y$ et $x \Leftarrow y$ est sémantiquement identique à l'équivalence $x \Leftrightarrow y$. Par exemple, sur la figure 5.2, les implications $A3 \Rightarrow B5$ et $A3 \Leftarrow B5$ seront comptabilisées comme 2 faux-positifs (parce qu'elles ne sont pas dans A_r) alors qu'elles sont sémantiquement identiques à l'équivalence

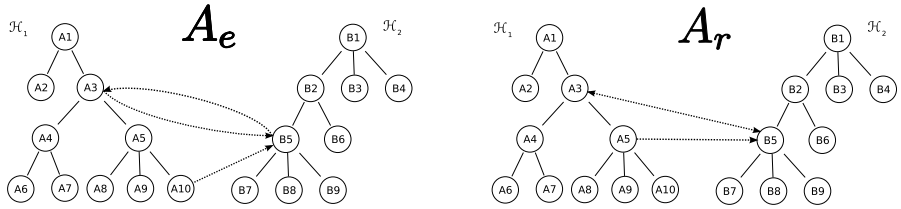


FIG. 5.2 – Deux alignements différents mais sémantiquement identiques

$A3 \Leftrightarrow B5$. Comme $A3 \Leftrightarrow B5$ ne figure pas explicitement dans A_e, elle sera comptabilisée comme 1 faux-négatif.

2. Le modèle ne prend pas en compte les déductions possibles et ne peut donc pas détecter les redondances. Sur la figure 5.2, l'implication $A10 \Rightarrow B5$ de V_e sera comptabilisée comme 1 faux-positif car elle ne figure pas dans A_r. Cependant, comme cette implication peut être déduite à partir de $A5 \Rightarrow B5$, elle ne doit pas être considérée comme un faux-positif. De manière opposée, l'implication $A5 \Rightarrow B5$ contenue dans A_r sera comptabilisé comme 1 faux-négatif car elle n'est pas dans A_e. Cependant, comme cette implication peut être déduite à partir de l'implication $A3 \Rightarrow B5$ de A_e, elle ne doit pas être considérée comme un faux-négatif.

A partir de l'exemple de la figure 5.2, nous obtenons des valeurs de rappel et de précision égales à 0 alors que les deux alignements sont sémantiquement identiques. Il est donc nécessaire de développer, dans le cas d'alignements considérant également l'implication, un modèle étendu prenant en compte les cas énoncés ci-dessus.

5.2 Modèle d'évaluation sémantique

Dans [Euz07], J. Euzenat propose des mesures sémantiques de précision et de rappel, afin de prendre en compte la sémantique de l'alignement et des modèles d'ontologie. Dans ce modèle sémantique, la fermeture de chaque ensemble de correspondances est également prise en compte.

Les nouvelles contingences de ce modèle sont données par la table 5.2. Avec ce nouveau modèle, les cardinalités des ensembles des vrai-positifs, faux-positifs, et faux-négatifs seront supérieures ou égales à celles du modèle classique. Seule la cardinalité des vrai-négatifs sera inférieure ou égale à celle du modèle classique.

5.2.1 Mesures sémantiques d'évaluation

Les mesures sémantiques de J. Euzenat ne prennent plus seulement en compte la présence ou l'absence d'éléments de correspondance d'un ensemble dans l'autre mais également la capacité de déduire les éléments de correspondances de l'autre ensemble à partir du premier et de la sémantique attachée aux ontologies comparées. Ces mesures sont appelées précision (P_s) et rappel

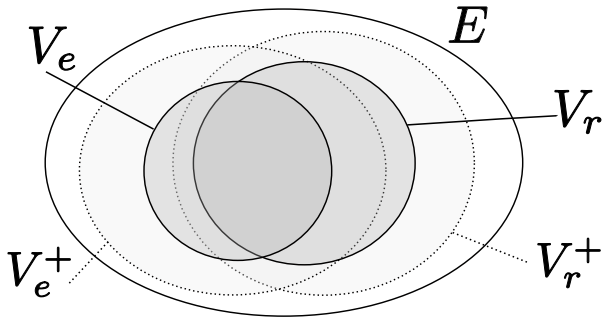


FIG. 5.3 – Diagramme du modèle d'évaluation sémantique

	pertinents	non pertinents	
trouvés	$\|V_e^+ \cap V_r^+\|$ vrai-positifs	$\|V_e^+ - V_r^+\|$ faux-positifs	$\|V_e^+\|$
non trouvés	$\|V_r^+ - V_e^+\|$ faux-négatifs	$\|(E - V_r^+) - V_e^+\|$ vrai-négatifs	$\|E - V_e^+\|$
	$\|V_r^+\|$	$\|E - V_r^+\|$	

TAB. 5.2 – Contingence des ensembles V_e^+ et V_r^+

(R_s) sémantiques. Leurs définitions, restreintes à notre modèle de hiérarchies introduit dans la section 2.1, sont données par :

$$P_s(V_e, V_r) = \frac{|V_e \cap V_r^+|}{|V_e|}$$

$$R_s(V_e, V_r) = \frac{|V_e^+ \cap V_r|}{|V_r|}$$

Contrairement aux mesures de précision et de rappel classiques, elles ne mesurent plus le taux de vrai-positifs par rapport aux nombres d'éléments trouvés ou pertinents. La précision sémantique évalue le nombre d'éléments de correspondances de V_e pouvant être déduits à partir de V_r (c.-à-d. le nombre d'éléments de V_e contenus dans la fermeture V_r^+) par rapport au nombre d'éléments de V_e. Le rappel sémantique évalue, quant à lui, le nombre d'élements de correspondance de V_r (de référence) pouvant être déduit à partir de V_e (c.-à-d. le nombre d'éléments de V_r contenus dans la fermeture V_e^+) par rapport au nombre d'éléments de V_r.

Comparaison des ordonnancements. Puisque d'une part, $|V_e \cap V_r^+| > |V_e \cap V_r|$ et que d'autre part $|V_e^+ \cap V_r| > |V_e \cap V_r|$, les mesures de précision et de rappel sémantiques seront toujours supérieures ou égales à celles de la précision et du rappel classiques.

5.2.2 Limites du modèle

Ces mesures reposent toujours, en partie, sur les ensembles V_e ou V_r. Ainsi, elles peuvent donner des résultats différents lorsque l'on considère deux alignements A_{e1} et A_{e2} identiques (c.-à-d. possédant la même fermeture $V_{e1}^+ = V_{e2}^+$) mais ayant des ensembles de correspondances différents $V_{e1} \neq V_{e2}$.

Exemple. La figure 5.4 présente trois alignements à évaluer A_{e1}, A_{e2} et A_{e3}, et un alignement de référence A_r. Les deux alignements A_{e1} et A_{e2} sont égaux puisque leurs fermetures sont égales : le seul élément de correspondance qui les différencie, $A10 \Rightarrow B5$, est déduit à partir de $A3 \Rightarrow B5$. Pourtant ces deux alignements n'obtiennent pas les même valeurs de précision sémantique par rapport à l'alignement de référence A_r :

$$P_s(V_{e1}, V_r) = \frac{|\{A3 \Rightarrow B5, A4 \Leftrightarrow B3, A10 \Rightarrow B5\}|}{|\{A3 \Rightarrow B5, A4 \Leftrightarrow B3, A10 \Rightarrow B5, A9 \Rightarrow B8\}|} = \frac{3}{4} = 0,75$$

$$P_s(V_{e2}, V_r) = \frac{|\{A3 \Rightarrow B5, A4 \Leftrightarrow B3\}|}{|\{A3 \Rightarrow B5, A4 \Leftrightarrow B3, A9 \Rightarrow B8\}|} = \frac{2}{3} = 0,67$$

L'alignement A_{e3} obtient quant à lui la valeur de précision sémantique suivante :

$$P_s(V_{e3}, V_r) = \frac{|\{A3 \Rightarrow B5, A4 \Leftrightarrow B3|}{|\{A3 \Rightarrow B5, A4 \Leftrightarrow B3, A6 \Leftrightarrow B4\}|} = \frac{2}{3} = 0,67$$

On remarque qu'en suivant ce principe, la mesure de précision sémantique peut être artificiellement augmentée. En effet, si l'on ajoute à l'alignement toutes les redondances d'un vrai-positif alors la précision sémantique augmente. Par contre, si cet élément de correspondance est faux-positif, le fait de rajouter ses redondances fait diminuer la précision. Afin d'obtenir un meilleur score de précision sémantique, on pourra ajouter toutes les redondances des éléments de correspondances qui ont obtenu les meilleurs scores de qualité et qui sont donc a priori sûrs d'être valides.

La mesure de rappel sémantique peut également faire l'objet des remarques ci-dessus. Si l'on considère deux alignements de référence égaux (à partir de leur fermeture) mais ne contenant pas les mêmes éléments, un même alignement à évaluer peut, dans ce cas, obtenir une valeur de rappel différente sur chacun des alignements de référence. Si l'on ajoute de la redondance dans l'alignement de référence, la cardinalité $|V_r|$ augmentera, et par conséquent la valeur du rappel sémantique diminuera.

Ces mesures prennent en compte de la même manière l'équivalence et l'implication. En effet, si un élément de correspondance est faux-positif (c.-à-d. non déductible à partir de l'alignement de référence), il entraînera une baisse de la précision sémantique identique quelque soit sa nature (implication ou équivalence). Cependant, comme une équivalence est une double implication, la baisse de précision engendrée par un faux-positif de type équivalence doit être supérieure à la baisse engendrée par un faux-positif de type implication.

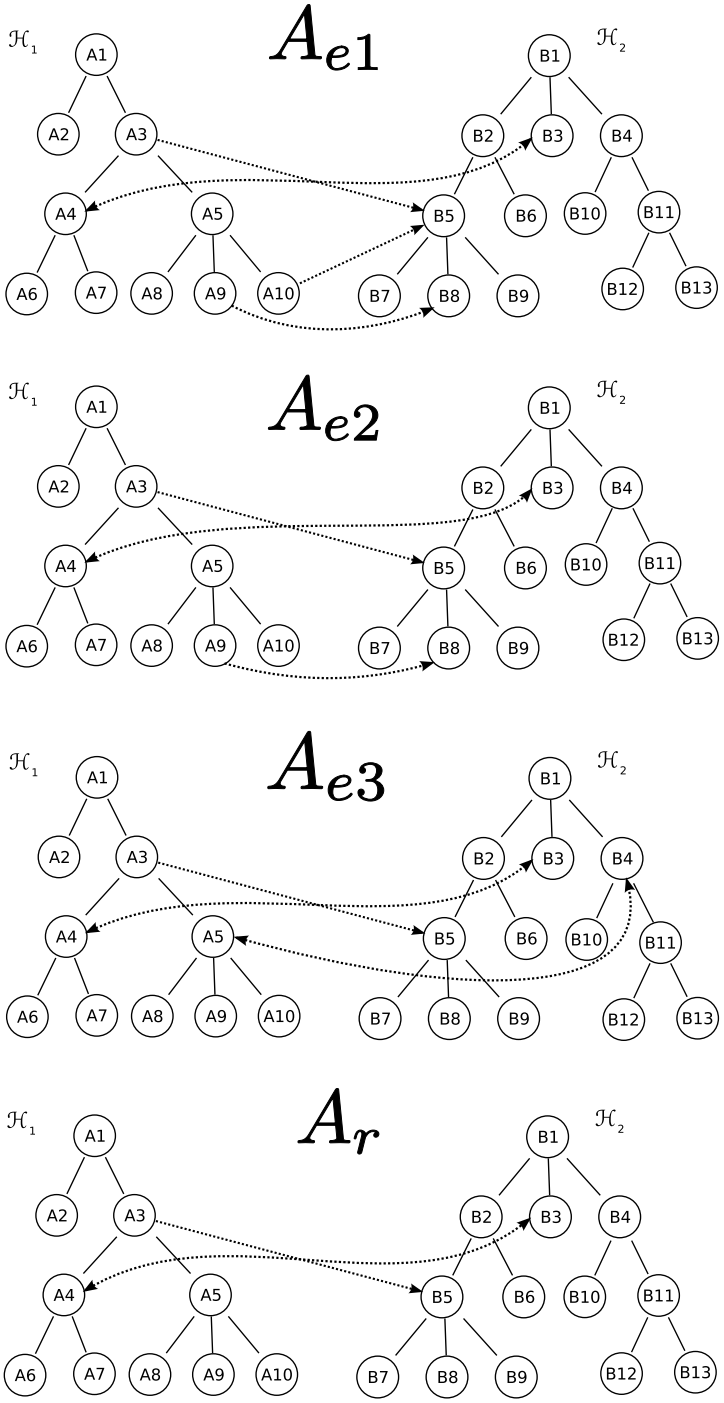


FIG. 5.4 – Trois alignements à évaluer et un alignement de référence

5.3 Adaptation du modèle de comparaison

Afin de pallier les limites du modèle classique, nous proposons, dans notre cadre d'alignement de hiérarchies, de nous appuyer sur un modèle d'alignement ne contenant que des implications. A partir d'un alignement $A = (V, q)$, les éléments de correspondance $x \Leftrightarrow y$ contenus dans V de type équivalence seront décomposés en deux implications $x \Rightarrow y$ et $x \Leftarrow y$. Ensuite afin d'éviter de quantifier positivement ou négativement la redondance, nous proposons d'utiliser les mesures de précision et rappel idéales, proposées dans [Euz07]. Ces mesures comparent l'alignement à évaluer et l'alignement de référence à partir de leurs fermetures transitives respectives.

$$P_i(V_e, V_r) = \frac{|V_e^+ \cap V_r^+|}{|V_e^+|}$$

$$R_i(V_e, V_r) = \frac{|V_e^+ \cap V_r^+|}{|V_r^+|}$$

Exemple. En reprenant les alignements de la figure 5.4, nous pouvons calculer les fermetures (voir déf. 2.8) de chaque alignement.

$$\begin{aligned} V_{e1}^+ = V_{e2}^+ = \big\{ & \{A3, A4, A5, A6, A7, A8, A9, A10\} \Rightarrow \{B1, B2, B5\}, \\ & \{A4, A6, A7\} \Rightarrow \{B3\}, A9 \Rightarrow B8, \{A4, A3, A1\} \Leftarrow \{B3\} \big\} \end{aligned}$$

$$\begin{aligned} V_{e3}^+ = \big\{ & \{A3, A4, A5, A6, A7, A8, A9, A10\} \Rightarrow \{B1, B2, B5\}, \\ & \{A4, A6, A7\} \Rightarrow \{B3\}, \{A5, A8, A9, A10\} \Rightarrow \{B4\}, \\ & \{A4, A3, A1\} \Leftarrow \{B3\}, \{A5, A3, A1\} \Leftarrow \{B4, B10, B11, B12, B13\} \big\} \end{aligned}$$

$$\begin{aligned} V_r^+ = \big\{ & \{A3, A4, A5, A6, A7, A8, A9, A10\} \Rightarrow \{B1, B2, B5\}, \\ & \{A4, A6, A7\} \Rightarrow \{B3\}, \{A4, A3, A1\} \Leftarrow \{B3\} \big\} \end{aligned}$$

A partir de ces fermetures, les valeurs des mesures de précision et rappel idéales sont données par :

$$\begin{aligned} P_i(V_{e1}, V_r) = P_i(V_{e2}, V_r) &= \frac{30}{31} = 0,97 \\ P_i(V_{e3}, V_r) &= \frac{30}{49} = 0,61 \\ R_i(V_{e1}, V_r) = R_i(V_{e2}, V_r) = R_i(V_{e3}, V_r) &= \frac{30}{30} = 1 \end{aligned}$$

Les ensembles de correspondance V_{e1} et V_{e2} sont mieux évalués par la mesure de précision idéale que par les mesures de précision sémantique et de précision

classique. Cependant, pour l'ensemble V_{e3}, c'est l'inverse qui se produit : $P_i < P_s \leq P$. Ce phénonème est expliqué par la capacité de la précision idéale à prendre en compte, pour chaque élément de correspondance, sa spécificité dans l'alignement. La spécificité d'un élément de correspondance $x \Rightarrow y$ dans un alignement $A = (V, q)$ est la différence entre la fermeture de V et celle de V privée de $x \Rightarrow y$:

$$\textit{spécificité}(x \Rightarrow y, V) = V^+ - [V - \{x \Rightarrow y\}]^+$$

Remarque. On peut noter qu'un élément de correspondance redondant dans un alignement aura une spécificité égale à l'ensemble vide. En effet, un élément de correspondance a est redondant si $(V - \{a\})^+ = V^+$ (voir section 2.2.3).

Par exemple, sur l'alignement A_{e2}, la spécificité de $A9 \Rightarrow B8$, seul l'élément de correspondance qui n'est pas dans V_r, est $\textit{spécificité}(A9 \Rightarrow B8) = \{A9 \Rightarrow B8\}$. L'ensemble des faux positifs spécifiquement engendrés par $A9 \Rightarrow B8$, représentant la partie de sa spécificité qui n'est pas dans la fermeture de V_r, est $V_r^+ - \textit{spécificité}(A9 \Rightarrow B8) = \{A9 \Rightarrow B8\}$. Sur l'ensemble de correspondance V_{e3}, la spécificité de $A5 \Leftrightarrow B4$, seul l'élément de V_{e3} qui n'est pas dans V_r, est $\textit{spécificité}(A5 \Leftrightarrow B4) = \{\{A5, A8, A9, A10\} \Rightarrow \{B4\}, \{A1, A3, A5\} \Leftarrow \{B4, B10, B11, B12, B13\}\}$. L'ensemble des faux positifs spécifiquement engendrés par $A5 \Leftrightarrow B4$ est $V_r^+ - \textit{spécificité}(A5 \Leftrightarrow B4) = \{\{A5, A8, A9, A10\} \Rightarrow \{B4\}, \{A1, A3, A5\} \Leftarrow \{B4, B10, B11, B12, B13\}\}$.

Nous remarquons que les faux-positifs spécifiques apportés par l'élément de correspondance $A5 \Leftrightarrow B4$ sont plus nombreux que ceux apportés par $A9 \Rightarrow B8$. Etant donné que les intersections $V_{e2} \cap V_r$ et $V_{e3} \cap V_r$ sont égales, la précision idéale $P_i(V_{e3}, V_r)$ sera donc plus faible que $P_i(V_{e2}, V_r)$.

Conclusion

Nous avons montré dans ce chapitre que le modèle d'évaluation classique des méthodes d'alignement, basé sur les mesures de précision et de rappel, n'est pas du tout adapté aux alignements implicatifs. Partant du modèle d'évaluation sémantique proposé par J. Euzenat [Euz07] et de la notion de fermeture présentée dans notre modèle d'alignement, nous proposons d'utiliser les mesures de précision et de rappel idéales. Pour cela, les équivalences des deux ensembles de correspondance à comparer sont décomposés en double implication. Ces mesures ont l'avantage de prendre en compte les capacités déductives de notre modèle d'alignement et permettent ainsi d'évaluer de la même manière deux alignements sémantiquement égaux.

Réalisations et évaluations expérimentales

6

Sommaire

Introduction

Ce chapitre présente tout d'abord les réalisations logicielles de la méthode AROMA et d'un prototype d'aide à visualisation, à la validation et à l'édition d'alignements. Ensuite, nous exposons les évaluations qui ont été menées dans le but d'étudier le comportement et la performance d'AROMA. Ces évaluations concernent d'une part, la phase de pré-traitement, et plus particulièrement la sélection des termes représentatifs, et d'autre part, l'extraction de l'alignement. Sur cette dernière série d'évaluations, nous avons utilisé deux jeux de tests. Le premier est composé de hiérarchies textuelles et les second est un ensemble d'ontologies RDFS/OWL.

6.1 Réalisations logicielles

Dans le cadre de notre recherche, nous avons développé les modèles de hiérarchie et d'alignement, ainsi que les algorithmes de notre méthode AROMA. Les composants développés permettent de charger, prétraiter et d'aligner des hiérarchies textuelles et des ontologies. Les hiérarchies textuelles sont décrites soit sous forme de fichier XML, soit disponibles dans une sous-arborescence d'un système de fichiers. Le format d'entrée pris en compte pour les ontologies est la langage RDFS/OWL.

Nous avons également intégré ces composants dans une application d'aide à l'alignement. Cette application permet non seulement de charger, de configurer les algorithmes et d'effectuer l'alignement, mais également de visualiser et de valider les alignements obtenus. Nous avons ainsi développé une visualisation des hiérarchies et leur alignement sous forme de graphe. Nous proposons également une série de filtres permettant à l'utilisateur d'alléger la représentation et de se concentrer sur certaines parties de l'alignement.

La réalisation de la visualisation a été effectuée, par Xavier Aimé, dans le cadre d'un mémoire d'ingénieur CNAM [Aim07].

6.1.1 Réalisation de la méthode AROMA

Nous avons réalisé la méthode AROMA dans le langage JAVA. L'outil comprend les modules suivants :

- le module de chargement de hiérarchies textuelles et d'ontologies au format RDFS/OWL,
- le module d'acquisition des termes,
- l'algorithme de sélection des termes pertinents,
- la représentation des hiérarchies sous le modèle introduit 2.1,
- les algorithmes d'extraction de règles entre hiérarchies,
- l'algorithme d'alignement syntaxique,
- les différents filtres (déduction des équivalences, élimination des inconsistances, filtre de réduction des cardinalités).

Concernant le chargement des ontologies RDFS/OWL, nous nous sommes appuyés sur la bibliothèque Jena[1] proposée par les laboratoires HP [McB02]. Pour le module d'acquisition des termes, nous avons utilisé des outils du traitement du langage existants. L'étiquetage morpho-syntaxique des textes est réalisé par l'étiqueteur de Stanford ([TM00], [TKMS03]). Cet étiqueteur, basé sur le modèle de maximisation de l'entropie, obtient de bonnes précision : 96,86% globalement (et 86,91% pour les mots précédemment non-rencontrés) sur le corpus Penn Treebank [MMS93]. Nous avons choisi cet étiqueteur parce qu'il est écrit en JAVA (et donc facilement intégrable à nos modules) et parce que le texte est étiqueté au format Penn Treebank qui est le format utilisé en entrée de l'extracteur de termes. La lemmatisation est faite par l'analyseur morphologique Morpha[2] [MCP01]. Pour l'extraction des termes binaires, nous utilisons l'outil ACABIT de l'équipe TAL du LINA[3] ([Dai94], [Dai03]). Le logiciel est écrit dans le langage PERL. Il prend en entrée un fichier dans un format pseudo XML, structurant le corpus en textes, titres etc. Le texte étiqueté doit être au format PennTreeBank. Nous avons également réalisé un extracteur de termes simples, permettant la sélection de termes de type « NOM ».

Afin d'implémenter les différentes mesures statistiques utilisées par la méthode AROMA, nous nous sommes appuyés sur la bibliothèque Commons-Math[4] de la fondation Apache. Pour la méthode d'alignement syntaxique, nous avons utilisé les mesures implémentées dans la bibliothèque SecondString[5] [CRF03] développée à l'université de Carnegie Mellon.

6.1.2 AROMAViz, un outil de visualisation et de validation interactive d'alignement

La méthode AROMA peut être utilisée dans un cadre d'alignement automatique de hiérarchies. Cependant, il est souvent nécessaire de faire valider un alignement avant de l'utiliser. Ainsi, nous avons proposé un outil permettant d'effectuer le calcul d'un alignement par la méthode AROMA, mais également une interface de visualisation d'alignement. Comme les hiérarchies alignées sont souvent de taille importante, la visualisation d'un alignement devient problématique car la représentation devient rapidement illisible. C'est pourquoi, nous avons également proposé une série de filtres permettant d'alléger la visualisation afin que l'utilisateur puisse se concentrer sur certaines parties de la hiérarchie ou certains types de relations de correspondance.

Les fonctionnalités mises en place dans l'outil sont :
- Le chargement des hiérarchies. Ces hiérarchies peuvent être, soit des ontologies au format RDFS/OWL, soit des hiérarchies textuelles décrites dans un fichier XML,
- La configuration et le calcul de l'alignement,
- La visualisation et l'édition de l'alignement.

[1] http ://jena.sourceforge.net
[2] http ://www.informatics.susx.ac.uk/research/groups/nlp/carroll/morph.html
[3] Equipe Traitement Automatique du Langage, Laboratoire Informatique de Nantes Atlantique
[4] http ://commons.apache.org/math
[5] http ://secondstring.sourceforge.net

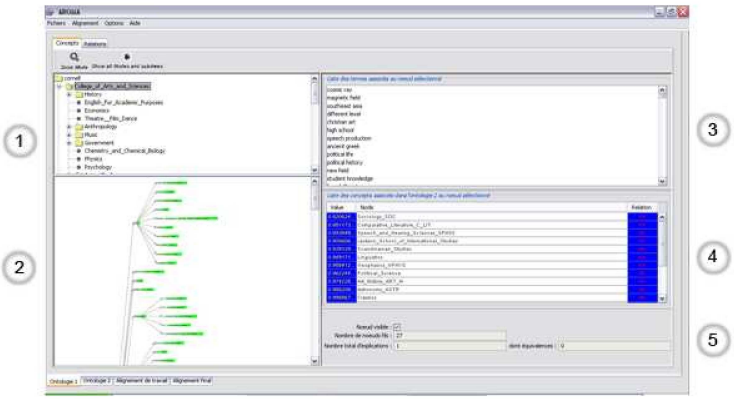


FIG. 6.1 – Espace hiérarchie

Concernant la visualisation et l'aide à la validation de l'alignement, l'IHM d'AROMAViz est centrée autour de deux types d'espaces de travail :

- Espace hiérarchie : cet espace permet de visualiser et d'éditer les informations sur l'alignement à partir de l'une des hiérarchies ;
- Espace alignement : cet espace permet de visualiser et d'éditer l'alignement de manière globale.

Espace hiérarchie

L'espace hiérarchie permet l'édition des informations sur une hiérarchie et les relations qui en sont issues. Cet espace est illustré sur la figure 6.1. Cet espace est découpé en deux grandes parties. Sur la partie gauche, l'utilisateur dispose de deux représentations de la hiérarchie courante. A partir de l'une de ces représentations, il peut sélectionner une entité afin de visualiser et d'éditer, sur la partie droite, certaines informations qui lui sont relatives.

Sur la figure 6.1, nous avons numéroté 5 zones : les deux premières concernent les représentations de la hiérarchie (partie de gauche) et les trois autres (partie de droite), la visualisation des informations relatives à une entité. Les 5 zones de l'espace hiérarchie sont les suivantes :

1. TreeView de la hiérarchie : elle représente la hiérarchie dans sa totalité. Cette vue est invariante des sélections que l'utilisateur peut choisir.
2. Aperçu du graphe de la hiérarchie : cette vue, sous forme de graphe, représente également la hiérarchie. Contrairement à la première vue de la hiérarchie, cette dernière répercute les différentes sélections que l'utilisateur a pu faire.
3. La liste des termes associés à l'entité sélectionnée.
4. La liste des éléments de correspondances dont fait partie l'entité. Pour chaque élément de correspondance, on affiche la valeur de qualité associée, l'entité qui est en correspondance avec l'entité courante, et la relation entretenue par les deux entités.
5. Informations complémentaires sur l'entité : les informations disponibles

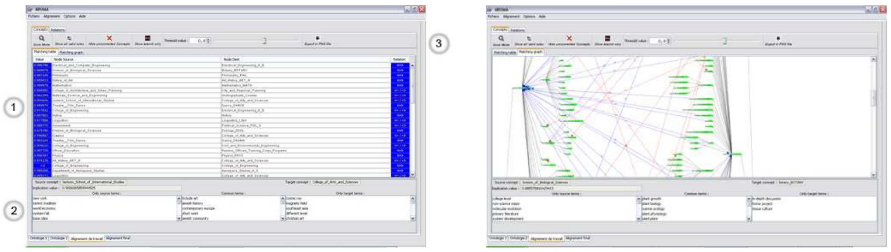


FIG. 6.2 – L'espace alignement

sont le nombre d'éléments de correspondance, ventilé par type (équivalence ou implication), dont l'entité courante fait partie et le nombre de descendants directs de l'entité.

Lors de la sélection d'une entité, un filtre permet également de rendre invisible le noeud courant ou de le sélectionner comme nouvelle racine. Ces modifications sont répercutées dans la vue 2 et dans la vue de l'alignement qui sera décrite après.

Espace alignement

L'espace alignement, illustré figure 6.2, est composé de 3 zones :

1. Zone de visualisation de l'alignement.
2. Zone de présentation d'informations relatives à un élément de correspondance.
3. Barre d'outils agissant sur la visualisation de l'alignement.

La zone de visualisation de l'alignement peut avoir deux apparences. La première (illustrée sur l'image de gauche, figure 6.2) permet de visualiser l'alignement sous forme de tableau. Chaque élément de correspondance est représenté par une ligne sur laquelle sont indiquées sa valeur de qualité, les deux entités et la relation qui le composent. Les lignes du tableau peuvent être triées par ordre de valeurs de qualité croissant ou décroissant, par ordre alphabétique des entités sources ou cibles, ou encore par type de relation. La deuxième apparence (illustrée sur l'image de droite, figure 6.2) présente l'alignement sous forme de graphe. Ce graphe représente, à chaque extrémité (gauche et droite), les deux hiérarchies dessinées de manière horizontale. Un élément de correspondance est schématisé par un arc reliant une entité de la hiérarchie de gauche et celle de droite. L'arc est valué par la valeur de qualité de l'élément qu'il représente et le type de la relation est donné par la couleur de l'arc (rouge pour l'équivalence et bleu pour l'implication).

A partir de la zone de visualisation, l'utilisateur peut sélectionner un élément de correspondance et visualiser ses informations détaillées dans la deuxième zone. Cette zone, illustrée figure 6.3, présente les noms des deux entités en correspondance, la valeur de qualité de l'élément de correspondance, ainsi que trois listes de termes. La liste de gauche présente les termes associés uniquement à l'entité source, la liste du milieu donne les termes associés aux deux entités, et celle de droite, les termes uniquement associés à l'entité cible. Dans le cas d'une

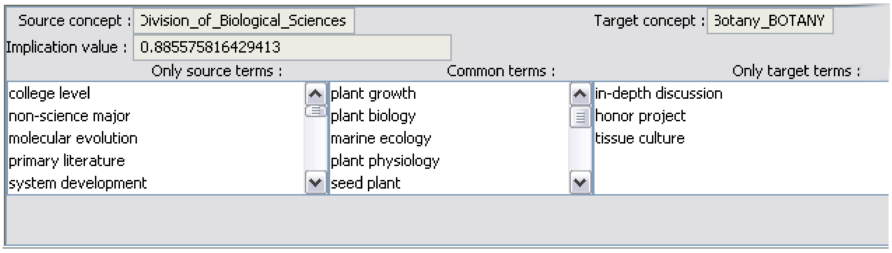


FIG. 6.3 – Présentation des informations relative à un élément de correspondance

implication, les termes de la liste de gauche sont les contre-exemples. Dans le cas d'une équivalence, les termes des listes de gauche et de droite sont les contre-exemples. Dans les deux cas, les termes de la liste du milieu sont les exemples de la relation considérée.

La barre d'outils donne accès aux différentes fonctionnalités de manipulation de la visualisation de l'alignement. Cette barre d'outils contient, tout d'abord, des fonctionnalités classiques telles que le zoom sur la représentation sous forme de graphe ou l'exportation en format image de la représentation graphique de l'alignement. Nous proposons également quelques fonctionnalités de filtrage permettant de rendre la visualisation graphique de l'alignement plus lisible et donc intelligible par l'utilisateur. Les filtres proposés sont :

- La sélection du seuil minimal de qualité à partir duquel les éléments de correspondance sont affichés. La valeur de ce seuil peut être comprise entre le seuil d'extraction (φ_r) et 1.
- La possibilité de masquer les éventuels éléments de correspondance redondants.
- La possibilité de masquer les entités n'intervenant dans aucun élément de correspondance.
- La restriction de l'affichage à l'une des deux composantes asymétriques : $A_{\mathcal{H}_1 \Rightarrow \mathcal{H}_2}$ ou $A_{\mathcal{H}_2 \Rightarrow \mathcal{H}_1}$.
- La restriction de l'affichage à la composante fortement symétrique (c.-à-d. l'affichage uniquement des éléments de correspondance de type équivalence).

La figure 6.4 illustre l'effet combiné du masquage des entités non alignées et de la variation du seuil minimal d'affichage des éléments de correspondance. La première représentation contient le graphe complet, puis avec les entités masquées, ensuite avec la modification de la valeur seuil. Cet exemple montre que la combinaison de différents filtres d'affichage permet d'alléger grandement la visualisation de hiérarchies. Ces filtres deviennent très utiles lorsque le nombre d'entités composant les hiérarchies est élevé.

Finalement, un dernier filtre permet de restreindre l'affichage à une branche de la hiérarchie. En sélectionnant une entité, l'utilisateur peut visualiser seulement les éléments de correspondances qui sont issus d'une de ses entités ascendantes ou descendantes. Ce filtre est illustré sur la figure 6.5. La branche sélectionnée apparaît en haut de la figure. Les entités cibles figurant dans les éléments de correspondance ainsi sélectionnés ne sont pas organisées selon leur structure hiérarchique.

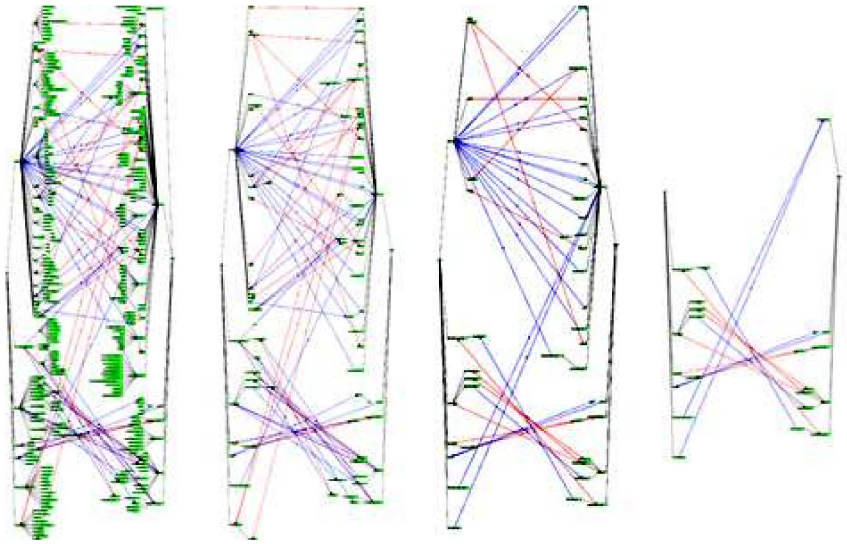


FIG. 6.4 – Exemple d'utilisation des filtres

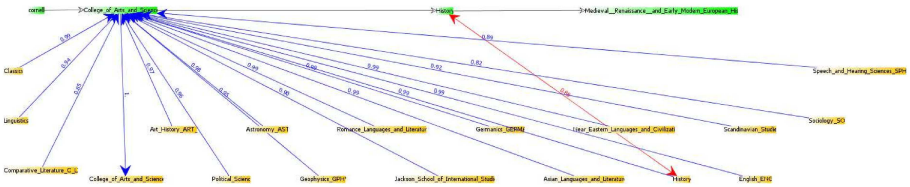


FIG. 6.5 – Exemple du filtre de sélection d'une branche

Sur chaque type de visualisation de l'alignement, l'utilisateur peut, lors de la sélection d'un élément de correspondance, le supprimer, l'annoter ou encore le valider. Dans le dernier cas, cet élément de correspondance sera ajouté dans un alignement final.

6.2 Démarche expérimentale

6.2.1 Objectifs et tests réalisés

Les expérimentations que nous avons menées ont eu pour objectifs d'étudier le comportement et la performance de la méthode AROMA sur deux types de structures hiérarchiques : les hiérarchies textuelles et les ontologies (décrites en RDFS/OWL). Sur le premier type de données, nous utilisons la phase de réindexation permettant d'extraire des termes puis de les sélectionner et de les associer aux entités de la hiérarchie. Dans le cas d'ontologies OWL, cette phase de sélection n'est pas utilisée étant donné qu'elles contiennent relativement peu de données textuelles.

Ces expérimentations contiennent, tout d'abord, l'évaluation de la phase de sélection des termes, puis, dans un deuxième temps, l'analyse des résultats obtenus d'une part sur l'alignement de hiérarchies textuelles, et d'autre part sur l'alignement d'ontologies OWL. Dans les deux cas, nous disposons de jeux de tests décrits dans la section suivante.

Dans le cadre des évaluations d'alignements, nous avons testé 5 versions de la méthode :

1. **méthode simple** : extraction des règles + déduction des équivalences ;
2. **méthode avec élimination des inconsistances** : méthode simple + filtre d'élimination des inconsistances ;
3. **méthode avec similarité syntaxique** : méthode intermédiaire + méthode d'alignement syntaxique ;
4. **méthode complète** : méthode simple + élimination des inconsistance + méthode d'alignement syntaxique ;
5. **méthode complète avec réduction de la cardinalité** : méthode complète + filtre de réduction de la cardinalité.

En fonction des résultats obtenus, nous présentons dans chaque cas (hiérarchies textuelles ou ontologies OWL), seulement les versions qui apportent un changement significatif.

Nous avons également étudié le comportement et la performance d'AROMA obtenus avec différentes mesures d'intérêt utilisées dans le contexte d'évaluation des règles d'association. Pour cela, nous avons sélectionné 6 mesures à partir de la classification proposée par [Bla05] et présentée section 1.2.2. La grande majorité de ces 6 mesures ont une valeur bornée entre 0 et 1 (seul l'indice de Loevinger peut avoir une valeur inférieure à 0). En effet, contrairement à la fouille de règles d'association où les mesures sont traditionnellement utilisées en post-traitement et dans le but d'ordonner les règles, nous les utilisons dans notre cas dans la phase d'extraction. Il est ainsi nécessaire que ces mesures soient bornées afin de permettre de choisir plus facilement le seuil de sélection.

Mesure	**Portée**	**Nature**	**Sujet**	**Valeur fixe**	**Définition**
II	$\Rightarrow$	S	I	$0,5$	$P(n_{a\bar{b}} < Poisson(\frac{n_a \cdot n_{\bar{b}}}{n}))$
Loevinger	$\Rightarrow$	D	I	0	$1 - \frac{n \cdot n_{a\bar{b}}}{n_a \cdot n_{\bar{b}}}$
IPEE	$\rightarrow$	S	E	$0,5$	$P(n_{a\bar{b}} < Binomiale(n_a, 1/2))$
Confiance	$\rightarrow$	D	E	$0,5$	n_{ab}/n_a
AVL	$\leftrightarrow$	S	I	$0,5$	$P(n_{ab} > Poisson(\frac{n_a \cdot n_b}{n}))$

TAB. 6.1 – Définitions et propriétés des mesures sélectionnées

Les 6 mesures sélectionnées sont la confiance, l'indice de Loevinger, l'intensité d'implication (II), l'indice probabiliste d'écart à l'équilibre (ipee), l'indice de la vraisemblance du lien (AVL), et l'indice de Jaccard. La table 6.1 montre pour chacune de ces mesures, sa portée (règle ($\rightarrow$), quasi-implication ($\Rightarrow$), quasi-conjonction ($\leftrightarrow$)), sa nature (statistique (S) ou descriptive (D)), son sujet (écart à l'indépendance (I) ou écart à l'équilibre (E)), la valeur fixe prise à l'indépendance ou à l'équilibre (fonction du sujet de la mesure) et sa définition.

6.2.2 Jeux de tests

Hiérarchies textuelles

Nous avons utilisé, pour les expérimentations, deux couples de hiérarchies textuelles. Ces deux jeux de tests ont été proposés dans [DMDH04].

Catalogues de cours. Ce jeu de tests est composé de deux catalogues de cours proposés par les universités de Cornell et Washington. Les descriptions textuelles des cours sont organisées de manière hiérarchique en écoles et collèges et ensuite en départements et centres. Les deux hiérarchies sont assez larges dans le sens où elle contiennent respectivement 166 et 176 entités organisées en 3 niveaux maximum. Ces hiérarchies sont respectivement associées à 4360 et 6957 descriptions de cours. Les descriptions des cours sont assez sommaires : elles contiennent le code identifiant du cours, son titre, des informations sur la date, les crédits, les enseignants du cours et éventuellement une petite description du contenu. Un exemple de description est donné par la figure 6.6. Le jeu de tests est également fourni avec un alignement contenant 54 éléments de correspondance entre les entités des catalogues Cornell et Washington. Cet alignement de référence n'est constitué que d'équivalences, et il est fonctionnel dans le sens Cornell vers Washington, c.-à-d. qu'une entité de Cornell sera associée au plus à une entité Washington, cependant, une entité de Washington peut être associée à plusieurs entités de Cornell.

Profils d'entreprises. Le jeu de tests "Company Profile" est issu des annuaires Web Yahoo.com et TheStandard.com. Ces hiérarchies contiennent respectivement 115 et 333 entités ainsi que 13634 et 9504 descriptions d'entreprises. Les descriptions d'entreprises sont organisées en secteurs puis en industries. Ces

EAS 101 Introductory Geological Sciences (i) Fall, spring, or summer. 3 credits. Fall, A. Moore ; spring, J. M. Bird ; summer, W. Brice.
Designed to enhance an appreciation of the physical world. Emphasizes natural environments, surface temperatures, and dynamic processes such as mountain belts, volcanoes, earthquakes, glaciers, and river systems. Interactions of the atmosphere, hydrosphere, biosphere, and lithosphere (earth system science). Water, mineral, and fuel resources ; environmental concerns. Field trips in the Ithaca region.

FIG. 6.6 – Exemple d'une description de cours de la hiérarchie Cornell

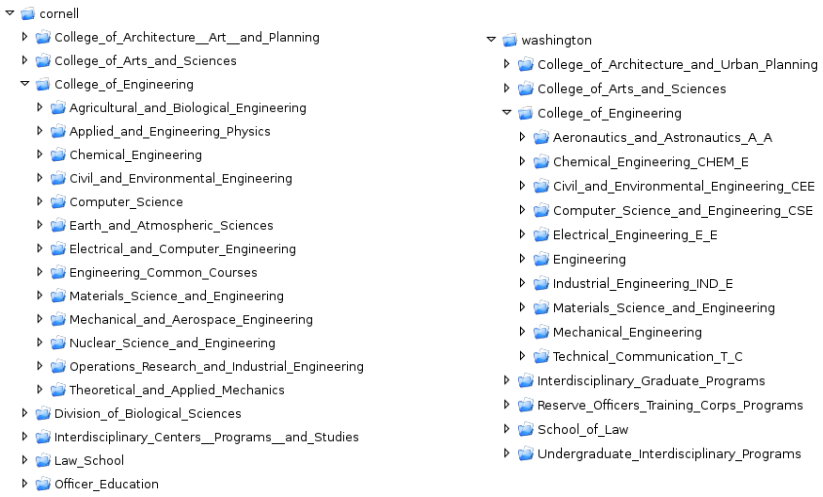


FIG. 6.7 – Extraits des structures des hiérarchies Cornell et Washington

Profile - Apple Computer, Inc. (NasdaqNM :AAPL)
Business Summary
Apple Computer, Inc. designs, manufactures and markets personal computers and related personal computing and communicating solutions for sale primarily to education, creative, consumer, and business customers. Substantially all of the Company's net sales to date have been derived from the sale of its Apple Macintosh line of personal computers and related software and peripherals. Apple Macintosh personal computers are characterized by their intuitive ease of use, innovative industrial designs and applications base, and built-in networking, graphics, and multimedia capabilities. The Company offers a range of personal computing products, including personal computers, related peripherals, software, and networking and connectivity products. All of the Company's Macintosh products employ PowerPC RISC-based microprocessors. More from Market Guide : Expanded Business Description
Financial Summary
AAPL designs, manufactures and markets personal computers and related personal computing and communicating solutions for the sale primarily to education, creative, consumer, and business customers. For the nine months ended 6/30/01, revenues fell 36% to $3.91 billion. Net loss before acct. change totaled $103 million, vs. an income of $616 million. Revenues reflect a decline in worldwide demand for the Company's products. Net loss reflects rebate programs and price cuts.

FIG. 6.8 – Exemple d'une description d'entreprise de la hiérarchie Yahoo

deux hiérarchies sont encore plus larges que les précédentes. La hiérarchie Standard, qui possède pratiquement 3 fois plus d'entités que Yahoo, couvre les mêmes domaines mais possède un découpage des activités d'entreprises beaucoup fin que celui de Yahoo.com.

Sur ce jeu de tests, nous ne disposons pas d'alignement de référence.

Ontologies

Jeu de tests de l'INRIA. Ce jeu de tests est utilisé dans le cadre de l'OAEI[6]. Ce jeu de tests contient une ontologie de référence sur les références bibliographiques et un ensemble d'ontologies de tests qui seront comparées avec la référence (la version du benchmark utilisée contient 50 tests). Pour chaque test, un alignement de référence est proposé. L'ontologie de référence contient 33 classes nommées, 24 propriétés inter-classes (*owl :ObjectProperty*), 40 propriétés simples (*owl :DatatypeProperty*), 56 instances nommées, et 20 instances anonymes. La majorité des ontologies tests (séries 1xx et 2xx) ont été générées de manière systématique à partir de la référence.

Les ontologies de test sont réparties en trois groupes :
- 1xx - Tests simples (4 tests) : comparaison de l'ontologie référence avec elle même, avec une ontologie non pertinente, et avec la restriction OWL-Lite et la généralisation OWL-Full de l'ontologie de référence.
- 2xx - Tests systématiques (42 tests) : comparaison avec des ontologies dégradées obtenues à partir de la référence.
- 3xx - Tests réels (4 tests) : comparaison avec 4 ontologies réelles sur les

[6]Ontology Alignment Evaluation Initiative est une initiative internationale visant à évaluer et à améliorer l'évaluation des techniques d'alignement d'ontologies,
http ://oaei.ontologymatching.org

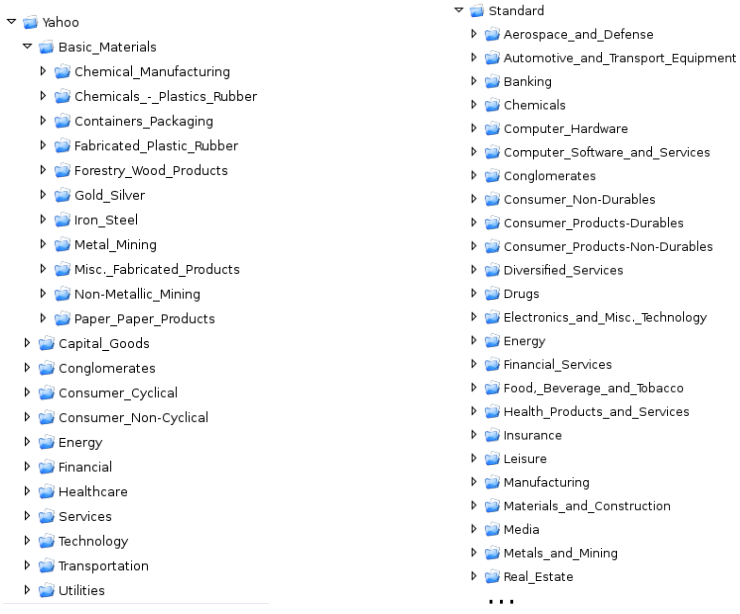


FIG. 6.9 – Extraits des structures des hiérarchies Yahoo et Standard

références bibliographiques.

Les dégradations effectuées sur les ontologies de la série 2xx portent sur les labels (suppression, synonymie, traduction, chaînes aléatoires), sur les commentaires (suppression ou traduction), sur l'ordre partiel (suppression, extension, restriction), sur les instances (suppression), sur les propriétés (suppression ou restrictions) ou encore sur les classes (extension ou restriction de leur définition).

6.3 Evaluation de la sélection des termes

La méthode de prétraitement des hiérarchies textuelles permet de redéfinir la relation qui, à chaque entité, lui associe un ensemble de documents, en une relation lui associant un ensemble de termes. Pour cela, cette méthode réalise, tout d'abord, l'extraction des termes contenus dans les documents, puis elle permet de les associer aux entités en les sélectionnant par l'évaluation des règles *terme* $\rightarrow$ *entité*. L'évaluation des règles est effectuée par l'utilisation de mesures de qualité. Nous avons opté pour l'utilisation de l'intensité d'implication, mais d'autres mesures peuvent être utilisables. Nous proposons, ainsi, dans cette section, d'évaluer différentes mesures pour la sélection des termes représentatifs des entités. Nous avons mené deux évaluations : une quantitative et une qualitative. L'évaluation quantitative a consisté à comptabiliser le nombre de termes sélectionnés en fonction de la mesure utilisée et du seuil de sélection choisi. Cette première évaluation nous a permis d'étudier le pouvoir filtrant des mesures dans ce contexte. L'évaluation qualitative a consisté à comparer les ensembles de termes extraits par les mesures ayant été retenues sur la première étude. Cette évaluation a eu pour objectif de mettre en évidence la complémentarité

des mesures et/ou leur ressemblance quant aux termes qu'elles sélectionnent.

6.3.1 Evaluation quantitative

Cette évaluation quantitative a pour objectif de montrer l'influence du seuil de sélection des termes sur la quantité de termes extraits. Nous avons mené cette évaluation sur les deux jeux de tests « Catalogues de cours » et « Profils d'entreprises ».

Nous avons sélectionné, pour cette évaluation, quatre mesures asymétriques et une mesure de similarité. Parmi les mesures asymétriques, nous avons choisi deux mesures statistiques (intensité d'implication et Ipee) et deux mesures descriptives (l'indice de Loevinger et la confiance). L'intensité d'implication et Loevinger sont des indices d'écart à l'indépendance, tandis que Ipee et la confiance sont des indices d'écart à l'équilibre. L'indice de similarité utilisé est celui de Jaccard.

Pour chacune de ces mesures, nous avons fait varier le seuil de sélection des termes φ_t de 0 à 1 (par pas de 0.02) et avons comptabilisé le nombre de termes sélectionnés. La figure 6.10 présente un graphique pour chacune des quatre hiérarchies étudiées. Chaque graphique contient cinq courbes (une par mesure) représentant le nombre de termes sélectionnés (en ordonnée) en fonction de la valeur du seuil de sélection des termes φ_t (en abscisse). L'origine des courbes (seuil de sélection égal à 0) donne le nombre de termes extraits avant sélection. Les hiérarchies Cornell et Washington contiennent beaucoup moins de termes (respectivement 35276 et 34719 termes) que les hiérarchies Yahoo et Standard (respectivement 126420 et 182396 termes).

Sur l'ensemble des graphiques, nous observons trois tendances. La première concerne l'évolution de la mesure de Jaccard. Avec cette mesure, le nombre de termes sélectionnés décroît très rapidement en fonction du seuil de sélection croissant. Nous remarquons également que cette décroissance est beaucoup plus rapide sur les hiérarchies qui contiennent le plus de termes. Cette mesure est ainsi très filtrante.

La deuxième tendance est celle des mesures asymétriques descriptives, l'indice de Loevinger et la confiance. Sur les deux courbes, la décroissance du nombre de termes sélectionnés est, contrairement à Jaccard, très lente. Les deux courbes présentent sur l'ensemble des graphiques une baisse soudaine (de 5% à 8%) lorsque le seuil de sélection dépasse la valeur de $0,5$. Ces courbes marquent plus marginalement une autre baisse un peu avant le seuil de $0,7$. Pour les deux indices et sur l'ensemble des hiérarchies, le nombre de termes sélectionnés reste élevé : avec un seuil de sélection égal à 1, 81% à 86% des termes initialement extraits sont sélectionnés.

La dernière tendance est celle des mesures probabilistes, l'intensité d'implication et Ipee. Les courbes de ces deux mesures comportent deux phases. Sur la première phase, le nombre de termes sélectionnés reste globalement stable, ensuite il chute très fortement et soudainement (baisse de 80% du nombre de termes sélectionnés pour une variation du seuil de $0,05$). Sur cette première phase, les courbes de Ipee marquent un léger décrochement (baisse de 5% à 7%) pour un seuil de $0,26$. Les courbes de l'intensité d'implication sont très stables.

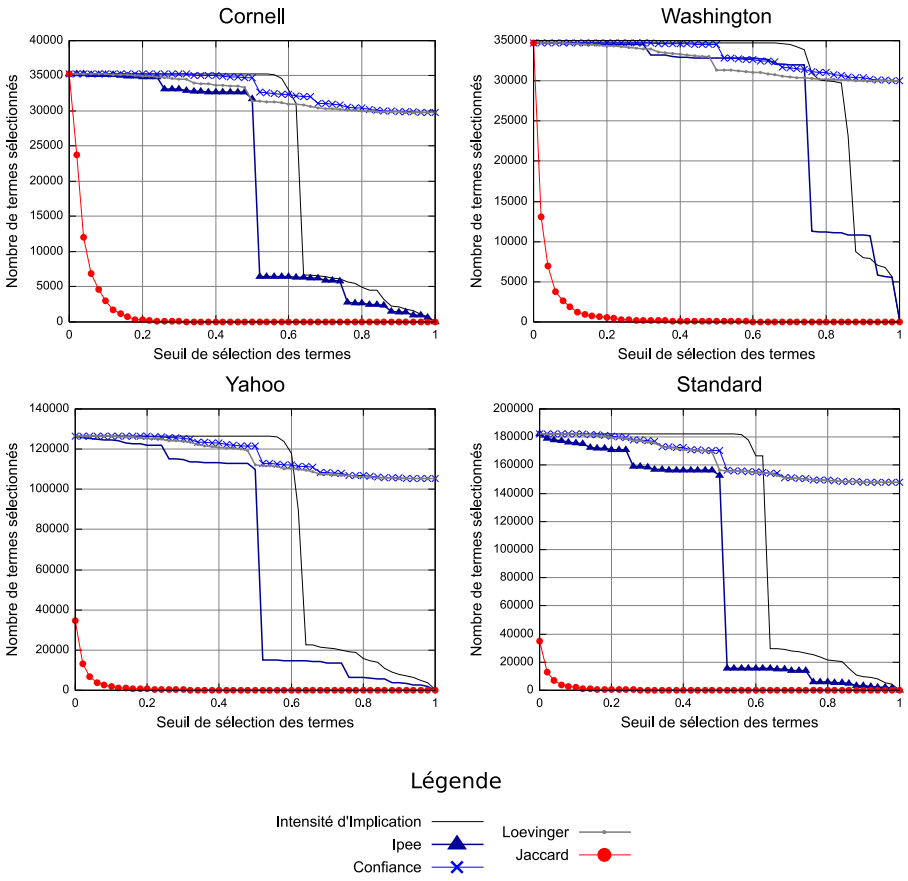


FIG. 6.10 – Evolution du nombre de termes sélectionnés en fonction de φ_t

Le décrochement brutal intervient, sur trois hiérarchies, à une valeur du seuil de $0,5$ pour Ipee et $0,6$ pour l'intensité d'implication. Sur la hiérarchie Washington, ces décrochements interviennent pour des valeurs de seuil plus élevées : respectivement $0,75$ et $0,87$. Sur la deuxième phase, les courbes baissent plus doucement jusqu'à un nombre de termes sélectionnés quasi nul (lorsque le seuil de sélection atteint 1). Les courbes de Ipee sont encore marquées par un plus léger décrochement (baisse de 30% à 35%) intervenant pour une valeur de seuil égale à $0,87$ pour les trois hiérarchies Cornell, Yahoo et Standard. Sur la hiérarchie Washingon, ce décrochement est un peu plus grand (baisse de 45%) et intervient pour une valeur de seuil plus élevée $(0,93)$. Pour l'intensité d'implication, ces décrochements sont moins marqués et interviennent pour des valeurs de seuil plus élevées.

Globalement, il apparaît qu'en fonction de leur nature, descriptive ou statistique, les mesures adoptent des comportements assez similaires quant au volume de termes sélectionnés. La mesure de similarité de Jaccard est très filtrante. Les mesures descriptives (l'indice de Loevinger et la confiance) ont, quant à elles, un pouvoir filtrant très faible. Les mesures statistiques (Ipee et l'intensité d'implication) ont un comportement binaire. Elles sont, dans un premier temps, pas filtrantes, puis deviennent très filtrantes à partir d'un seuil respectivement égal à $0,5$ et $0,6$. De par leur pouvoir filtrant adaptable, les mesures statistiques semblent être les mieux adaptées à la sélection des termes.

Mesure	Plage d'utilisation
Jaccard	$[0 - 0,15]$
Intensité d'Implication	$[0,6 - 1]$
Ipee	$[0,5 - 1]$

TAB. 6.2 – Plage d'utilisation des mesures

6.3.2 Etude qualitative

Les mesures statistiques, Ipee et l'intensité d'implication, ont un comportement similaire sur la quantité de termes extraits en fonction du seuil de sélection choisi. Afin de vérifier si elles ont également tendance à extraire les mêmes termes, nous avons comparé les ensembles de termes qu'elles ont respectivement sélectionnés et associés à chaque entité.

Le principe consiste à réaliser l'étape de sélection des termes avec chacune des deux mesures. Les ensembles de termes associés à chaque entité, respectivement par Ipee et par l'intensité d'implication, sont ensuite comparés. Nous avons utilisé pour cela la mesure de Dice. Nous avons répété cette opération pour chaque couple de valeur seuil possible en faisant varier le seuil de sélection φ_t de 0 à 1 (par pas de $0,02$). Pour chaque couple, nous avons gardé la moyenne des valeurs de similarité obtenues. Les résultats de cette expérimentation sont donnés sur la figure 6.11. Chaque graphique présente l'évolution des valeurs de la similarité de Dice en fonction des valeurs de seuils respectivement choisies pour Ipee et l'intensité d'implication. Parmi les quatre graphiques, trois sont très similaires. Celui de Washington diffère des autres dans le sens où les tendances

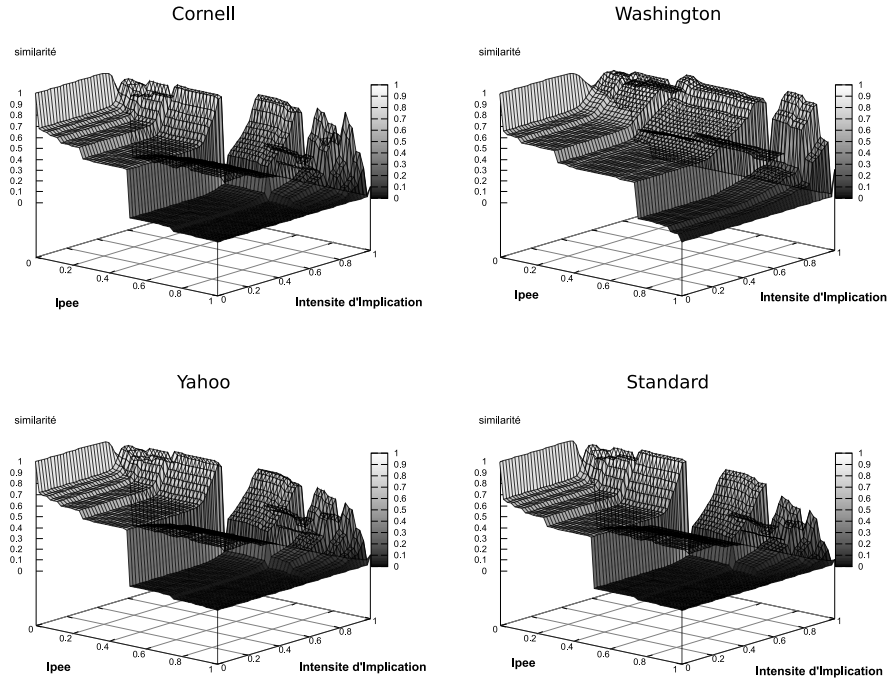


FIG. 6.11 – Evolution des similarités (de Dice) moyennes entre les ensembles de termes sélectionnés en fonction des seuils φ_t utilisés par une sélection s'appuyant sur les mesures Ipee et d'intensité d'implication

observées sont décalées vers des valeurs d'Ipee et d'intensité d'implication plus élevées.

Chaque paysage est découpé en deux vallées (zone composée de faibles valeurs données par la mesure de Dice) correspondant aux croisements d'intervalles de valeurs seuil incompatibles $[0,5;1] \times [0;0,6]$ et $[0;0,5] \times [0,6;1]$ (sur chaque croisement, le premier intervalle correspond à Ipee, le second à l'intensité d'implication). Une première arête (ligne représentant des fortes valeurs) est située au seuil 0 pour Ipee et sur l'intervalle $[0;0,6]$ pour l'intensité d'implication. Cette arête représente le cas où aucune des deux mesures n'est filtrante : dans ce cas, chaque entité est associée à son ensemble de termes initial. Une deuxième arête est située sur l'intervalle $[0,02;0,5]$ pour Ipee et pour une valeur d'intensité d'implication approximativement égale à $0,6$. Cette arête signifie que les ensembles de termes associés aux entités, obtenus pour un seuil d'intensité d'implication au voisinage de $0,6$ sont similaires à ceux obtenus par une sélection par Ipee avec un seuil compris entre $0,02$ et $0,5$. Une troisième arête est également présente sur l'intervalle $[0,5;0,75]$ d'Ipee et pour une valeur seuil d'intensité d'implication approximativement égale à $0,8$. Cette arête est cependant moins élevée que les deux premières (valeur de Dice égale à $0,8$). Finalement, sur le dernier croisement d'intervalles $[0,8;1] \times [0,9;1]$, deux pics moins élevés (ayant des valeurs de Dice respectives de $0,7$ et $0,5$ sur Cornell) sont présents. Lorsque

les valeurs de seuil approchent de 1 pour Ipee, la sélection des termes devient alors trop filtrante et les ensembles de termes associés aux entités deviennent ainsi très petits, voire vides.

En conclusion, les ensembles de termes sélectionnés par Ipee sur les plages de valeurs $[0,5; 0,8]$ (resp. $[0,8; 0,85]$) peuvent être approximés en utilisant l'intensité d'implication avec un seuil égal à $0,85$ (resp. $0,90$). Sur la dernière plage de valeurs $[0,85-1]$, la sélection d'Ipee devient très différente de celle de l'intensité d'implication. A part sur l'intervalle $[0; 0,5]$, où l'intensité d'implication est inadaptée (pouvoir de sélection des termes nul), Ipee ne permet pas d'approximer correctement les termes sélectionnés par cette première mesure.

6.3.3 Bilan

Sur ces expérimentations, nous avons exposé, tout d'abord, une étude quantitative de l'influence des mesures et des seuils de sélection sur le nombre de termes sélectionnés et associés aux entités des hiérarchies. Les mesures statistiques (Ipee et l'intensité d'implication) ont de meilleures capacités filtrantes que les mesures asymétriques descriptives. L'étude comparative des ensembles de termes sélectionnés par Ipee et l'intensité d'implication, montre que cette dernière mesure permet d'approximer dans nombreux cas, les résultats obtenus par Ipee.

6.4 Evaluation d'AROMA sur des hiérarchies textuelles

Cette section présente une étude de la performance de la méthode AROMA pour l'alignement de hiérarchies textuelles. Nous avons utilisé le jeu de tests « catalogues de cours ». Nous avons confronté les résultats produits par AROMA et l'alignement de référence fourni. L'alignement de référence est fonctionnel et est composé uniquement d'éléments de correspondance de type équivalence.

Sur ces tests, nous avons utilisé l'intensité d'implication pour la phase de sélection des termes. Concernant la phase d'extraction de l'alignement, nous avons comparé les résultats obtenus par les 6 mesures d'intérêt sélectionnées 6.2. Pour chacune de ces évaluations, nous avons fait varier le seuil de sélection des termes, φ_t, de $0,6$ à 1 et le seuil de sélection des règles de 0 à 1. Nous présentons, pour chaque version d'AROMA, les évolutions de la F-mesure en fonction des seuils de sélection choisis. Nous donnons également, dans chaque cas et pour chaque mesure, le meilleur score obtenu ainsi que le contexte c.-à-d. les valeurs des seuils.

Finalement, nous présentons les résultats obtenus en utilisant le modèle d'évaluation présenté section 5.3.

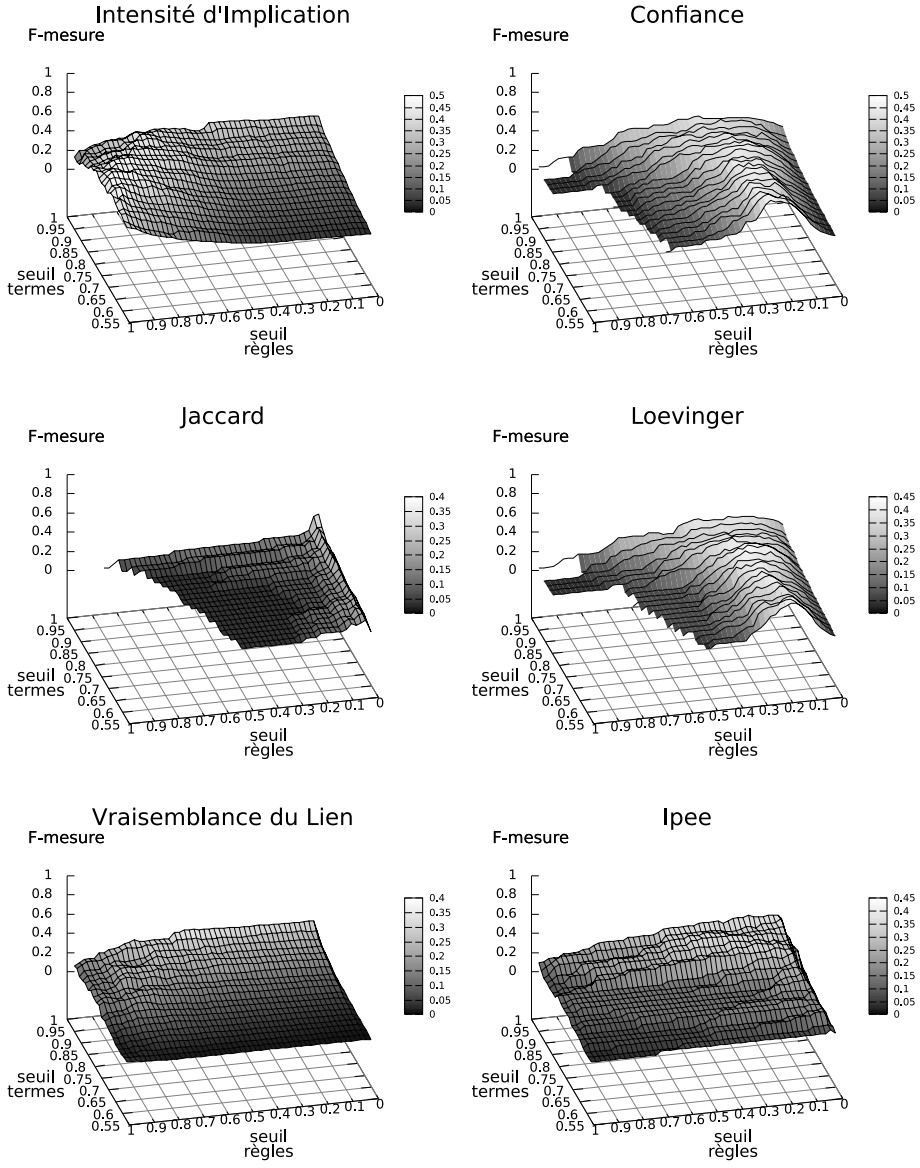


FIG. 6.12 – Evolution de la valeur de F-mesure, en fonction des seuils φ_t et φ_r, sur l'alignement de Cornell-Washington et en utilisant la méthode simple

6.4.1 Méthode simple

Sur cette première évaluation, les évolutions de la F-mesure présentent deux tendances globales. La première tendance est celle obtenue avec les mesures descriptives (la confiance, Loevinger et plus marginalement Jaccard). La deuxième tendance est celle des mesures statistiques (l'intensité d'implication, la vraisemblance du lien et plus marginalement Ipee).

Avec les mesures descriptives, l'influence de la valeur du seuil de sélection des règles est plus marquée que celle du seuil de sélection des termes. La F-mesure augmente rapidement en fonction du seuil de sélection des règles croissant pour atteindre un maximum assez rapidement (légèrement après $0,2$ pour la confiance et légèrement avant $0,2$ pour Loevinger), voire très rapidement dans le cas de la mesure de Jaccard (son maximum est atteint pour un seuil égal à $0,02$). Ensuite, la valeur de F-mesure redescend pour atteindre des valeurs nulles ou non définies. L'influence du seuil de sélection des termes se joue au niveau de l'étalement et l'aplatissement de l'évolution de la F-mesure. En effet, l'évolution des scores a tendance à être plus étalée et moins élevée lorsque le seuil de sélection des termes augmente. Cette tendance est aisément observable pour la confiance et Loevinger, mais elle est très atténuée pour la mesure de Jaccard.

Avec les mesures statistiques, on observe deux tendances bien distinctes : celles de l'intensité d'implication et de la vraisemblance du lien d'une part, et celle de Ipee d'autre part. Avec les deux premières mesures (intensité d'implication et vraisemblance du lien), les évolutions de F-mesures sont influencées par les valeurs des deux seuils de sélection. L'influence des seuils évolue en fonction de leurs valeurs. En effet, les surfaces représentant l'évolution de la F-mesure peuvent être découpées en deux parties : une première partie lorsque le seuil de sélection des règles est compris entre 0 et $0,5$, et une deuxième partie lorsque ce seuil est compris entre $0,5$ et 1. Sur la première partie ($\varphi_r \in [0; 0,5]$), la valeur de F- mesure est croissante en fonction de la valeur du seuil de sélection des termes croissante. Sur la deuxième partie, l'évolution est assez particulière. Lorsque le seuil de sélection des termes est inférieur à $0,85$, la valeur de F-mesure est croissante en fonction de la valeur du seuil de sélection des règles croissante. Pour une valeur de sélection des termes supérieure à $0,85$, la F-mesure est décroissante en fonction des valeurs de seuils croissantes.

Ipee a une évolution complètement différente des autres mesures statistiques. Son évolution peut s'apparenter quelque peu à celle de Jaccard. En effet, les deux mesures obtiennent leurs meilleurs scores pour des valeurs du seuil de sélection des règles très proches de 0. Ensuite, elles ont une évolution de F-mesure décroissante en fonction de la valeur du seuil de sélection des règles croissante. Cependant l'évolution des performances de Ipee est beaucoup plus douce que celle de Jaccard et l'effet de la valeur du seuil de sélection des termes est beaucoup plus marqué avec Ipee.

Au niveau performance, l'intensité d'implication obtient le meilleur score, en terme de F-mesure. Elle est suivie par la confiance et l'indice de Loevinger. Ensuite, Ipee obtient un score de $0,40$. Finalement, les mesures de similarité, la vraisemblance du lien et Jaccard, arrivent en dernière position et obtiennent à peu près les mêmes scores.

Mesure	φ_t	φ_r	**F-mesure**	**Précision**	**Rappel**
intensité d'implication	0,78	0,94	0,49	0,61	0,42
confiance	0,62	0,26	0,47	0,58	0,4
Jaccard	0,98	0,0	0,36	0,31	0,43
Loevinger	0,84	0,22	0,45	0,46	0,43
VdL	0,98	0,5	0,36	0,36	0,36
Ipee	0,98	0,14	0,4	0,55	0,32

TAB. 6.3 – Meilleures valeurs de F-mesure obtenues par chaque mesure sur l'alignement Cornell-Washington avec la méthode simple

Mesure	**Précision**			**Rappel**		
	moy.	é.-t.	méd.	moy.	é.-t.	méd.
intensité d'implication	0,24	0,19	0,18	0,48	0,11	0,51
confiance	0,55	0,31	0,54	0,23	0,23	0,15
Jaccard	0,83	0,28	1	0,04	0,09	0,02
Loevinger	0,50	0,28	0,46	0,19	0,23	0,08
VdL	0,1	0,11	0,06	0,33	0,14	0,36
Ipee	0,82	0,23	1	0,12	0,11	0,08

TAB. 6.4 – Statistiques sur les résultats obtenus par chaque mesure sur l'alignement Cornell-Washington avec la méthode simple

A part pour la confiance, ces meilleurs scores sont obtenus avec des valeurs du seuil de sélection des termes assez élevées. Au niveau du seuil de sélection des règles, seules les mesures asymétriques d'écart à l'indépendance obtiennent leur meilleur score avec une valeur de seuil de sélection en accord avec leur sémantique. En effet, le meilleur score pour l'intensité d'implication est obtenu avec une valeur de seuil φ_r supérieure à $0,5$ (valeur prise à l'indépendance par cette mesure), et pour l'indice de Loevinger, la valeur du seuil est supérieure à sa valeur prise à l'indépendance de 0.

A partir de la table 6.4, on remarque que les mesures de Jaccard et Ipee obtiennent des très bonnes précisions mais des rappels très faibles (surtout dans la cas de l'indice de Jaccard). La confiance et Loevinger ont des précisions moyennes et des rappels assez faibles. L'intensité d'implication obtient le meilleur rappel moyen mais sa précision moyenne est relativement faible. La vraisemblance du lien possède la plus mauvaise précision moyenne.

Au regard des scores moyens et médians obtenus par Jaccard et Ipee, on remarque que sur une grande plage de seuils, la méthode basée sur ces mesures ne trouve que très peu d'éléments de correspondance.

6.4.2 Méthode simple avec élimination des inconsistances

Dans ce cas, les formes d'évolutions de F-mesure sont les mêmes que celles obtenues avec la méthode simple (figure 6.12). L'élimination des inconsistances permet néanmoins d'augmenter significativement les valeurs de F-mesures maxi-

Mesure	φ_t	φ_r	**F-mesure**	**Précision**	**Rappel**
intensité d'implication	0,72	0,94	0,55	0,62	0,49
confiance	0,60	0,2	0,54	0,53	0,55
Jaccard	0,98	0,0	0,38	0,31	0,51
Loevinger	0,62	0,18	0,52	0,52	0,53
VdL	0,98	0,44	0,38	0,33	0,45
Ipee	0,98	0,14	0,44	0,56	0,36

TAB. 6.5 – Meilleures valeurs de F-mesure obtenues par chaque mesure sur l'alignement Cornell-Washington avec élimination des inconsistances

Mesure	**Précision**			**Rappel**		
	moy.	é.-t.	méd.	moy.	é.-t.	méd.
intensité d'implication	0,28	0,22	0,23	0,48	0,10	0,51
confiance	0,58	0,3	0,58	0,23	0,22	0,15
Jaccard	0,55	0,26	0,5	0,06	0,10	0,02
Loevinger	0,54	0,27	0,50	0,19	0,22	0,08
VdL	0,11	0,10	0,08	0,43	0,16	0,47
Ipee	0,63	0,14	0,67	0,13	0,11	0,09

TAB. 6.6 – Statistiques sur les résultats obtenus par chaque mesure sur l'alignement Cornell-Washington avec élimination des inconsistances

males obtenues par les mesures. La confiance et Loevinger sont les mesures qui profitent le mieux de cette augmentation (+7 points) suivies de près par les mesures d'intensité d'implication (+6 points). Ensuite Ipee bénéficie d'une hausse de 4 points. Les deux mesures de similarité gagnent seulement 2 points.

Ces meilleurs scores sont obtenus avec des valeurs de seuils assez proches ou égales à celles utilisées avec la méthode simple. Seul l'indice de Loevinger obtient son meilleur score avec une valeur du seuil de sélection des termes largement plus élevée $(+0,22)$.

Avec les mesures asymétriques d'écart à l'indépendance (intensité d'implication et Loevinger), l'élimination des inconsistances permet d'augmenter sensiblement leur précision moyenne. Par contre, cela procure l'effet inverse avec les mesures de Jaccard et Ipee qui voient leur précision moyenne baisser significativement. Dans le cas de la vraisemblance du lien, l'élimination des inconsistances lui permet d'améliorer son rappel moyen.

Globalement, malgré une amélioration significative des meilleurs résultats, le filtre d'élimination des inconsistances utilisé seul, n'apporte pas beaucoup de changements sur les résultats moyens de la plupart des mesures.

6.4.3 Méthode simple avec méthode syntaxique

Sur cette troisième évaluation, on peut remarquer que l'utilisation de la méthode d'alignement syntaxique permet d'améliorer notablement les résultats

Mesure	φ_t	φ_r	**F-mesure**	**Précision**	**Rappel**
intensité d'implication	0,92	0,96	0,75	0,88	0,66
confiance	0,62	0,26	0,67	0,64	0,7
Jaccard	0,94	0,18	0,68	0,78	0,6
Loevinger	0,64	0,26	0,69	0,69	0,68
VdL	0,96	0,86	0,67	0,73	0,62
Ipee	0,88	0,5	0,7	0,8	0,62

TAB. 6.7 – Meilleures valeurs de F-mesure obtenues par chaque mesure sur l'alignement Cornell-Washington avec la méthode syntaxique

par rapport à la méthode simple. Les mesures de similarité sont celles qui bénéficient le plus de cette amélioration : Jaccard et la vraisemblance du lien gagnent respectivement 32 et 31 points. Ensuite, ce sont les mesures asymétriques et statistiques qui profitent le mieux de cette hausse, suivies des deux mesures asymétriques descriptives. L'amélioration est globalement ressentie à la fois sur les valeurs de précision et de rappel.

Le classement des mesures a changé par rapport à la méthode simple. L'intensité d'implication obtient toujours le meilleurs score. Ipee, qui obtenait le plus mauvais score avec la méthode simple est désormais en seconde place. Les mesures de Loevinger et de Jaccard sont également passées au dessus de la confiance, qui obtient la dernière position avec la vraisemblance du lien.

A part pour la confiance, les valeurs du seuil de sélection des règles, avec lesquelles sont obtenus les meilleurs scores, sont assez différentes de celles utilisées avec la méthode simple ou avec le filtre d'élimination des inconsistances. La valeur du seuil de sélection des règles est plus élevée. Cette différence amène à penser que l'alignement syntaxique fonctionne mieux lorsque que la précision de l'alignement d'entrée est assez élevée.

Quant à la valeur du seuil de sélection des termes, l'intensité d'implication obtient son meilleur résultat avec une valeur beaucoup plus élevée que celles utilisées avec la méthode simple ou la méthode basée sur le filtre d'élimination des inconsistances. Les autres mesures ont tendance à avoir des valeurs du seuil de sélection des termes assez proches de celles utilisées avec la seconde méthode. Seule Ipee utilise une valeur de seuil largement moins élevée.

Pour les mesures qui obtenaient des valeurs de précision inférieures à $0,8$ avec la méthode simple, l'alignement syntaxique permet d'augmenter la précision moyenne. Dans les cas de l'intensité d'implication et de la vraisemblance du lien, le gain de précision entraîne un étalement des valeurs plus important que celui obtenu par la méthode simple. Dans le cas des mesures de Jaccard et Ipee, qui obtiennent des précisions moyennes supérieures à $0,8$ avec la méthode simple, la baisse de précision est assez conséquente (plus de dix points en moins). On peut toutefois remarquer que la méthode syntaxique permet de resserrer les valeurs de précision obtenues (baisse de l'écart-type resp. égale à $-0,13$ et $-0,07$).

La méthode syntaxique permet, pour l'ensemble des mesures, d'augmenter très significativement les valeurs de rappel et également de les resserrer autour de la moyenne. Ce gain de rappel est toutefois largement plus élevé pour les

Mesure	**Précision**			**Rappel**		
	moy.	é.-t.	méd.	moy.	é.-t.	méd.
intensité d'implication	0,28	0,22	0,21	0,63	0,06	0,64
confiance	0,57	0,22	0,70	0,60	0,05	0,58
Jaccard	0,71	0,11	0,74	0,57	0,02	0,58
Loevinger	0,59	0,21	0,70	0,59	0,05	0,57
VdL	0,16	0,17	0,09	0,47	0,11	0,51
Ipee	0,71	0,12	0,73	0,59	0,02	0,58

TAB. 6.8 – Statistiques sur les résultats obtenus par chaque mesure sur l'alignement Cornell-Washington avec la méthode syntaxique

Mesure	φ_t	φ_r	**F-mesure**	**Précision**	**Rappel**
intensité d'implication	0,94	0,96	0,76	0,86	0,68
confiance	0,6	0,26	0,73	0,75	0,72
Jaccard	0,94	0,18	0,72	0,85	0,62
Loevinger	0,64	0,28	0,71	0,78	0,66
VdL	0,98	0,98	0,70	0,80	0,62
Ipee	0,94	0,52	0,69	0,77	0,62

TAB. 6.9 – Meilleures valeurs de F-mesure obtenues par chaque mesure sur l'alignement Cornell-Washington avec la méthode complète

mesures qui obtiennent un mauvais rappel avec la méthode simple.

Sur ce jeu de tests, l'enrichissement de l'alignement par la méthode syntaxique permet d'améliorer significativement les résultats obtenus. Elle permet également de resserrer les différences entre les mesures que l'on observe dans le cas de la méthode simple.

6.4.4 Méthode complète

L'utilisation simultanée du filtre d'élimination des inconsistances et de la méthode d'alignement syntaxique permet d'obtenir des résultats sensiblement meilleurs que ceux obtenus par la seule utilisation de la méthode syntaxique. Seule Ipee obtient un score légèrement en baisse par rapport à celui obtenu avec l'utilisation de la méthode syntaxique. Le gain se joue principalement par l'augmentation des valeurs de précision obtenues (sauf pour l'intensité d'implication et Ipee). Le gain de rappel est très marginal voir nul.

Au niveau du classement, l'intensité d'implication obtient toujours le meilleur résultat, avec 3 points de plus que la confiance. Ensuite viennent les mesures de Jaccard, Loevinger, vraisemblance du lien et finalement Ipee qui retourne en dernière position.

Sur la table 6.10, on n'observe pas de grandes différences par rapport aux résultats obtenus avec la méthode d'alignement syntaxique (table 6.8). On remarque une légère amélioration des résultats en termes de précision pour les

Mesure	**Précision**			**Rappel**		
	moy.	é.-t.	méd.	moy.	é.-t.	méd.
intensité d'implication	0,32	0,21	0,27	0,63	0,07	0,64
confiance	0,59	0,21	0,69	0,60	0,04	0,58
Jaccard	0,67	0,13	0,70	0,52	0,07	0,53
Loevinger	0,61	0,19	0,70	0,58	0,04	0,57
VdL	0,15	0,16	0,09	0,56	0,11	0,60
Ipee	0,67	0,11	0,70	0,55	0,04	0,55

TAB. 6.10 – Statistiques sur les résultats obtenus par chaque mesure sur l'alignement Cornell-Washington avec la méthode complète

mesures asymétriques (mis à part Ipee) et une légère baisse pour les mesures de similarités. Au niveau du rappel, seule la vraisemblance du lien obtient une augmentation significative de son rappel moyen. Les rappels de Jaccard et Ipee ont quant à elles tendance à diminuer.

6.4.5 Méthode complète avec réduction de la cardinalité

En utilisant la connaissance du fait que l'alignement de référence est fonctionnel, l'utilisation du filtre de réduction de la cardinalité permet d'améliorer, dans certains cas, la performance de la méthode. En effet, les meilleurs résultats obtenus dans ce cas (présentés table 6.11) montrent une amélioration pour l'intensité d'implication, la confiance, l'indice de Loevinger, et la vraisemblance du lien. Le meilleur score d'Ipee ne s'améliore pas. Pour l'indice de Jaccard, le meilleur score est même en baisse. Pour la majorité des mesures, les valeurs utilisées pour obtenir ces meilleurs scores ne changent pas énormément par rapport à celles utilisées avec la méthode complète.

On peut également remarquer des différences assez notables entre les évolutions de la F-mesure de la méthode simple (figure 6.12) et celles que l'on obtient avec la méthode complète et le filtre de réduction des cardinalités (figure 6.13). Les mesures descriptives et asymétriques (la confiance et Loevinger) deviennent stables, après $\varphi_r = 0,2$, lorsque que la valeur du seuil de sélection des règles augmente, alors qu'avec la méthode simple, la F-mesure baissait très rapidement après une valeur seuil au environ de $0,2$. Avec ces deux mesures, une zone de F-mesure maximale est située au voisinage des valeurs de seuil φ_t et φ_r égales respectivement à $0,6$ et $0,2$.

L'évolution de la F-mesure obtenue, dans ce cas, avec la mesure de Jaccard est très différente de celle issue de la méthode simple. En effet, avec la première méthode la F-mesure décroît très rapidement en fonction du seuil de sélection de règles croissant. Sur cette dernière expérimentation son évolution est assez particulière : elle est en forme d'escalier et les bons scores sont obtenus soit avec une valeur du seuil de sélection des termes élevée, soit avec une valeur du seuil de sélection des règles élevée.

L'apparence des évolutions de F-mesure de l'intensité d'implication et de la vraisemblance du lien est assez comparable à celles de la méthode simple.

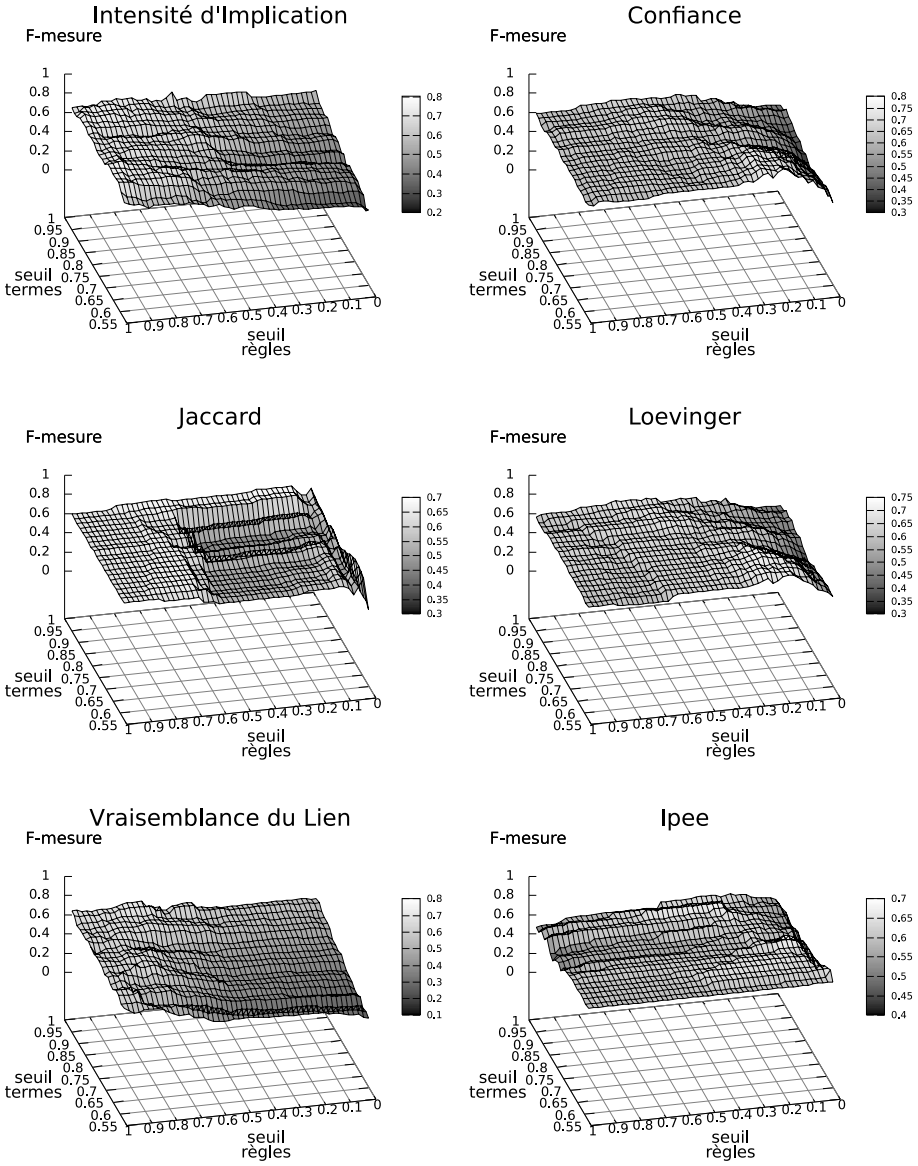


FIG. 6.13 – Evolution de la valeur de F-mesure, en fonction des seuils φ_t et φ_r, sur l'alignement de Cornell-Washington et en utilisant la méthode complète avec réduction de la cardinalité

Mesure	φ_t	φ_r	**F-mesure**	**Précision**	**Rappel**
intensité d'implication	0,92	0,96	0,77	0,88	0,68
confiance	0,6	0,2	0,77	0,78	0,75
Jaccard	0,94	0,1	0,69	0,79	0,62
Loevinger	0,68	0,24	0,73	0,76	0,7
VdL	0,98	0,7	0,71	0,76	0,66
Ipee	0,94	0,38	0,69	0,79	0,62

TAB. 6.11 – Meilleures valeurs de F-mesure obtenues par chaque mesure sur l'alignement Cornell-Washington avec la méthode complète + réduction de cardinalité

Mesure	**Précision**			**Rappel**		
	moy.	é.-t.	méd.	moy.	é.-t.	méd.
intensité d'implication	0,53	0,15	0,53	0,66	0,05	0,66
confiance	0,64	0,12	0,67	0,6	0,05	0,58
Jaccard	0,69	0,07	0,69	0,54	0,06	0,55
Loevinger	0,64	0,11	0,67	0,59	0,05	0,58
VdL	0,44	0,15	0,44	0,6	0,06	0,6
Ipee	0,69	0,06	0,7	0,55	0,03	0,55

TAB. 6.12 – Statistiques sur les résultats obtenus par chaque mesure sur l'alignement Cornell-Washington avec la méthode complète + réduction de cardinalité

La tendance de la F-mesure à décroître lorsque les valeurs des deux seuils de sélection tendent vers 1 disparaît avec cette dernière méthode.

La surface représentant l'évolution de la F-mesure obtenue avec Ipee devient presque plane alors qu'elle avait une tendance à décroître en fonction de la croissance de la valeur du seuil de sélection des règles sur la méthode simple.

La table 6.12 permet de confirmer l'amélioration des résultats obtenus sur la majorité des mesures. On remarque une nette amélioration de la précision moyenne pour l'intensité d'implication et de la vraisemblance du lien. Cette amélioration est plus marginale pour la confiance, l'indice de Loevinger et Ipee. Cependant, la précision moyenne de l'indice de Jaccard est en baisse.

L'utilisation du filtre de réduction cardinalité a également tendance à réduire l'effet des valeurs de seuil. En effet, les écart-types de précision ont tendance à diminuer significativement. Au niveau du rappel, on ne note pas de changement majeur par rapport aux valeurs obtenues sur la méthode complète.

En conclusion, lorsque que l'alignement recherché est fonctionnel, l'utilisation de cette connaissance permet d'améliorer de façon significative les résultats obtenus par AROMA.

Mesure	φ_t	φ_r	**F-mesure**	**Précision**	**Rappel**
intensité d'implication	0,72	0,98	0,78	0,82 (0,81)	0,75 (0,70)
confiance	0,98	0,68	0,69	0,70 (0,83)	0,68 (0,61)
Jaccard	0,98	0,7	0,72	0,83 (0,80)	0,63 (0,61)
Loevinger	0,98	0,5	0,68	0,68 (0,77)	0,68 (0,63)
VdL	0,98	0,76	0,75	0,83 (0,81)	0,68 (0,68)
Ipee	0,96	0,52	0,70	0,74 (0,78)	0,67 (0,59)

TAB. 6.13 – Meilleures valeurs de F-mesure (modèle sémantique idéal) obtenues par chaque mesure sur l'alignement Cornell-Washington avec la méthode complète + réduction de la cardinalité

6.4.6 Prise en compte des implications - évaluation avec mesures idéales

Sur cette dernière évaluation, nous avons utilisé la méthode complète avec réduction de la cardinalité. Nous avons procédé au calcul des mesures de précision et de rappel idéales avec le modèle introduit section 5.3. Dans ce cas, nous avons pris en compte non seulement les éléments de correspondance de type équivalence mais également ceux de type implication dans le sens Cornell vers Washington.

Comme pour les évaluations précédentes, nous présentons le tableau 6.13 des meilleurs scores calculés selon les mesures idéales de précision et rappel. Les scores apparaissant entre paranthèses sont ceux obtenus en considérant uniquement les relations d'équivalence.

Avec les mesures idéales de précision et de rappel, l'intensité obtient le meilleur score. La vraisemblance du lien, l'indice de Jaccard, et Ipee passent désormais devant la confiance et l'indice de Loevinger.

En comparant les résultats obtenus avec la prise en compte des implications et ceux obtenus avec uniquement les équivalences (ceux entre parenthèse), on remarque que les mesures statistiques et asymétriques (l'intensité d'implication et Ipee) sont celles qui profitent le mieux de l'aspect implicatif d'AROMA. En effet, l'intensité d'implication et Ipee présentent des augmentations respectives de $+0,01$ et $+0,04$ sur la précision et de $+0,05$ et $+0,08$ sur la rappel. Les similarités (indice de Jaccard et la vraisemblance du lien) ont des augmentations un peu moins élevées, notamment au niveau du rappel ($+0,02$ pour l'indice de Jaccard et 0 pour la vraisemblance du lien). La confiance et l'indice de Loevinger ont des valeurs de rappel qui augmentent significativement (resp. $+0,07$ et $+0,02$) mais en contrepartie, leur valeur de précision subit une baisse plus importante (resp. $-0,13$ et $-0,11$).

6.5 Evaluation d'AROMA sur des ontologies OWL

Comme pour l'évaluation d'AROMA sur les hiérarchies textuelles, nous avons testé les cinq versions de la méthode. Etant donné que le filtre d'élimination des inconsistances n'apporte pas de changement significatif sur les alignements produits, nous présentons seulement les résultats obtenus avec la méthode simple, avec l'utilisation de la similarité syntaxique et avec la méthode complète utilisant la réduction de la cardinalité. Dans chaque cas, nous présentons les courbes d'évolution de la F-mesure en fonction du seuil de sélection des règles. Ces courbes ont été réalisées pour les trois catégories du jeu de tests (1xx, 2xx et 3xx) et également pour l'ensemble du jeu de tests. La F-mesure a été calculée à partir des moyennes harmoniques de la précision et du rappel obtenus sur l'ensemble des alignements concernés. Finalement, nous comparons les résultats obtenus (avec la mesure qui obtient les meilleurs résultats) par rapport aux méthodes évaluées lors de la campagne OAEI 2006[7].

6.5.1 Méthode simple

Sur les trois premiers graphiques (mais également sur le graphique global), on observe globalement les mêmes formes d'évolution des courbes de F-mesure. Seules les valeurs maximales changent beaucoup entre les différentes séries de tests. Sans surprise, les meilleurs scores sont obtenus sur la série 1xx. Sur les séries 2xx et 3xx les mesures obtiennent de moins bons scores respectivement en baisse de $0,2$ et $0,4$ points.

Quant à la forme des courbes, on remarque que la confiance, l'indice de Loevinger, Ipee, et l'indice de Jaccard ont les mêmes tendances d'évolution : une très forte hausse au départ (entre 0 et $0,2$), puis une stabilisation voire une baisse vers des seuils élevés. Les courbes de la confiance et de l'indice de Loevinger sont pratiquement confondues. L'intensité d'implication obtient un peu près la même forme de courbe mais avec un décalage de $0,5$ au niveau du seuil. En effet, lorsque la valeur du seuil est comprise entre 0 et $0,5$, la courbe de l'intensité d'implication est stable. La courbe de la vraisemblance du lien est globalement stable jusqu'à une valeur seuil proche de $0,8$, ensuite les valeurs de F-mesure augmentent. On peut également remarquer que les courbes des mesures statistiques et asymétriques (l'intensité d'implication et Ipee) entament une décroissance lorsque le seuil de sélection des règles s'approche de 1.

En termes de performance, Ipee présente les meilleurs scores lorsque la valeur du seuil est inférieure à $0,5$. Ensuite, les mesures descriptives et asymétriques prennent le relais. Pour ces trois mesures, les valeurs optimales de F-mesure sont atteintes pour une valeur du seuil de sélection des règles égale à $0,5$. Les scores obtenus par l'intensité d'implication, puis respectivement par l'indice de Jaccard et la vraisemblance du lien sont globalement moins bons que ceux obtenus avec les trois premières mesures.

[7]Cette campagne d'évaluation des outils d'alignement d'ontologies a eu lieu lors du workshop « International Workshop on Ontology Matching » [SEN+06] qui s'est déroulé conjointement à la conférence ISWC 2006. Les résultats sont disponibles sur le site

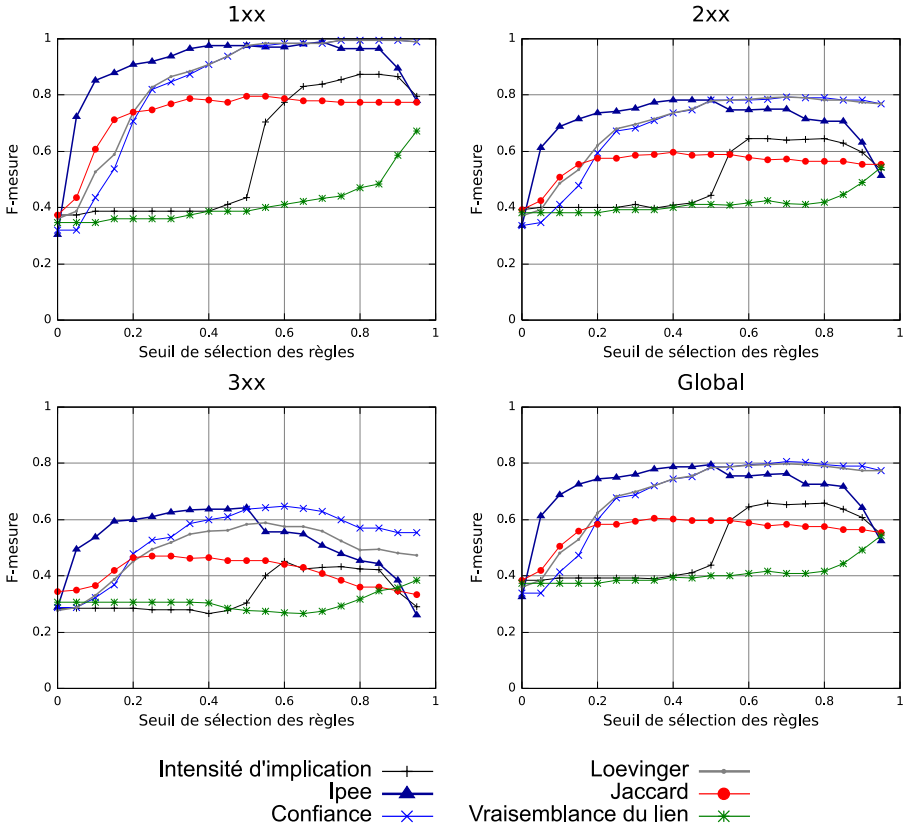


FIG. 6.14 – Evolutions des valeurs de F-mesure en fonction du seuil de sélection des règles sur la méthode simple

6.5.2 Méthode simple avec méthode syntaxique

L'effet de la méthode syntaxique se fait remarquer principalement sur les mesures qui obtenaient les moins bons scores sur l'évaluation précédente. L'indice de Jaccard semble être la mesure qui en profite le mieux. En effet, sa courbe de F-mesure a la même allure que celles de la confiance, de l'indice de Loevinger et d'Ipee. Son score moyen augmente d'environ $0,15$. Sur la plage $[0,6;1]$ (de valeurs du seuil), elle obtient désormais de meilleurs scores que l'intensité d'implication. L'intensité d'implication et la vraisemblance du lien bénéficient également de nettes améliorations, même si elles sont moins élevées que celle de l'indice de Jaccard.

Avec les trois meilleures mesures (la confiance, Loevinger et Ipee), les effets de la méthode syntaxique se font moins ressentir. On remarque tout de même une disparition de la tendance de décroissance, observée lorsque que la valeur du seuil tend vers 1. Cela affecte surtout Ipee et l'intensité d'implication.

Concernant les meilleurs scores, Ipee obtient toujours les meilleurs résultats sur la plage $[0;0,5]$. Cependant, sur la plage $[0,5;1]$, elle rivalise désormais (mis à part sur la série 2xx) avec la confiance et l'indice de Loevinger.

6.5.3 Méthode complète avec réduction de la cardinalité

Avec la méthode complète couplée avec le filtre de réduction de cardinalité, les courbes de résultats obtenues sont très différentes des précédentes. En effet, toutes les mesures ont désormais des courbes d'évolution de la F-mesure très similaires. Ces courbes deviennent très stables et au niveau des optima : la méthode ne semble plus être trop influencée par la valeur du seuil de sélection des règles. Ce phénomène, déjà en partie observé avec les hiérarchies textuelles, est dû à l'utilisation d'un critère de sélection local des éléments de correspondance. En effet, le filtre de réduction de cardinalité sélectionne, pour chaque entité de la hiérarchie source, l'élément de correspondance qui maximise le critère de qualité. On remarque toujours une légère décroissance lorsque le seuil de sélection des règles atteint la valeur de 1.

Sur la série de tests 3xx, les performances de l'indice de Jaccard semblent être un peu en retrait par rapport aux autres mesures. Sur la série 3xx, la vraisemblance du lien obtient également des résultats moins bons que les autre mesures (et même légèrement au- dessous de l'indice de jaccard). Sur la plage $[0;0,6]$, l'intensité d'implication obtient les mêmes résultats que la vraisemblance du lien, puis elle rejoint ensuite le groupe des meilleures mesures.

En conclusion, l'utilisation du filtre de réduction de la cardinalité (et donc d'un critère de sélection par maximisation locale), permet à AROMA d'obtenir des bons résultats sans avoir à se soucier du choix de la valeur du seuil. Cela est également en grande partie dû à la nature fonctionnelle de l'alignement de référence.

---

http ://oaei.ontologymatching.org/2006/

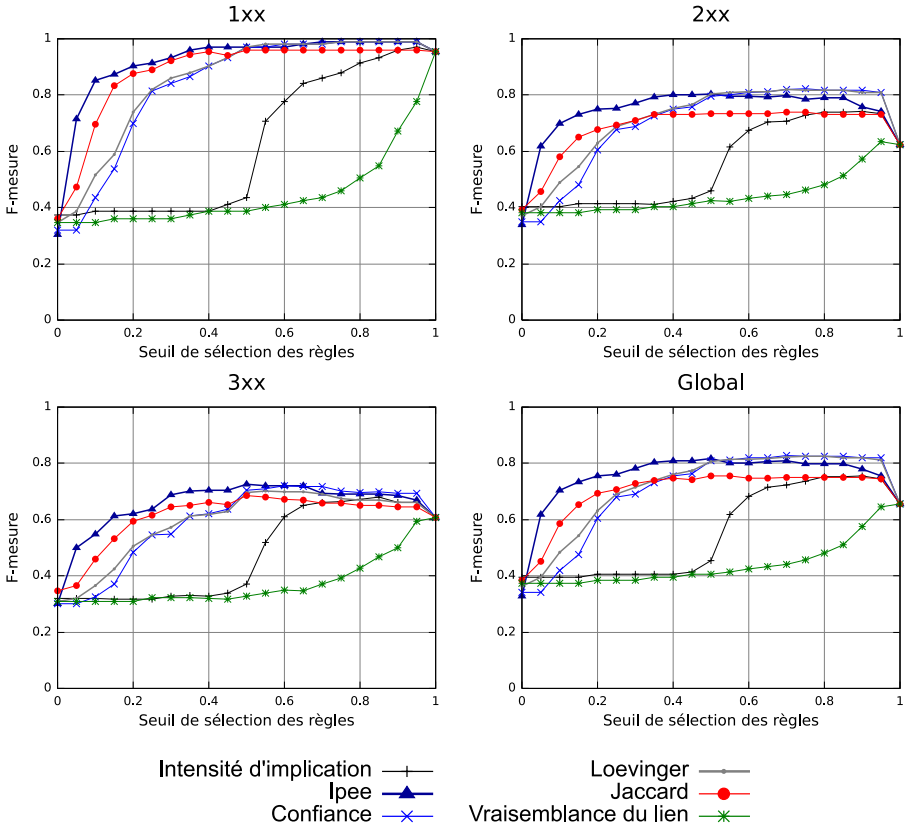


FIG. 6.15 – Evolutions des valeurs de F-mesure en fonction du seuil de sélection des règles sur la méthode avec alignement syntaxique

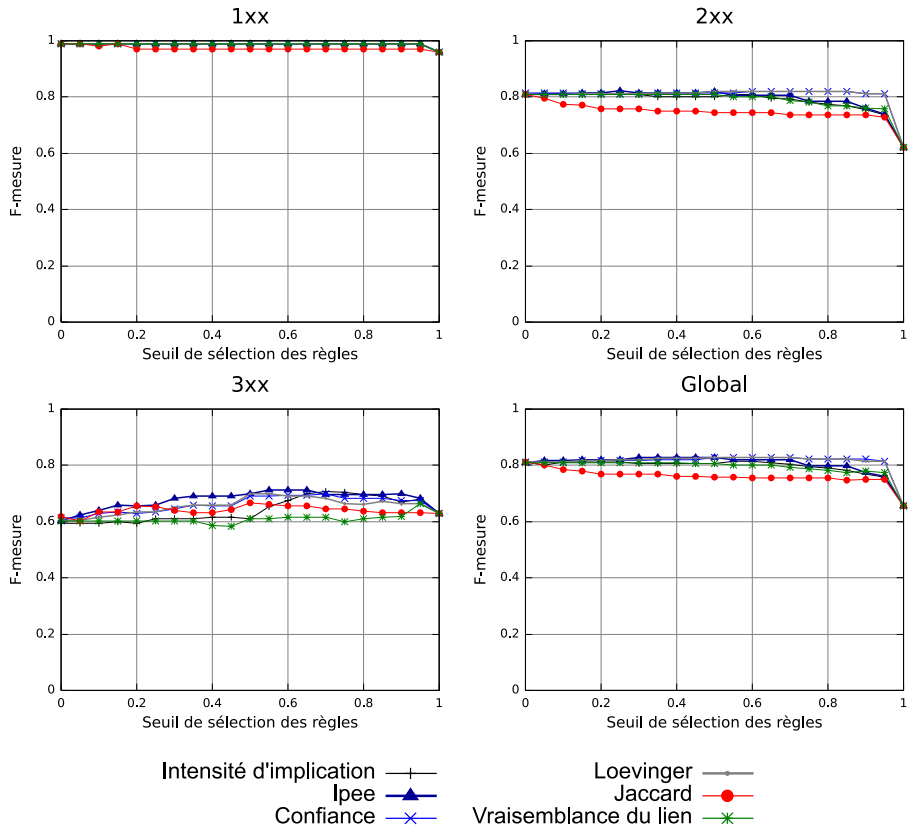


FIG. 6.16 – Evolutions des valeurs de F-mesure en fonction du seuil de sélection des règles sur méthode complète avec réduction de la cardinalité

algo	**edna**		**automs**		**coma**		**DSSim**		**falcon**	
test	P	R	P	R	P	R	P	R	P	R
1xx	0,96	1,00	0,94	1,00	1,00	1,00	1,00	0,98	1,00	1,00
2xx	0,90	0,49	0,94	0,64	0,96	0,82	0,99	0,49	0,91	0,85
3xx	0,94	0,61	0,91	0,70	0,84	0,69	0,90	0,78	0,89	0,78
Moy.	0,91	0,54	0,94	0,67	0,96	0,83	0,98	0,55	0,92	0,86

algo	**hmatch**		**jhuapl**		**OCM**		**prior**		**RiMOM**		**AROMA**	
test	P	R	P	R	P	R	P	R	P	R	P	R
1xx	0,91	1,00	1,00	1,00	0,95	1,00	1,00	1,00	1,00	1,00	0.99	0.99
2xx	0,83	0,51	0,20	0,86	0,93	0,51	0,95	0,58	0,97	0,87	0.95	0.70
3xx	0,78	0,57	0,18	0,50	0,89	0,51	0,85	0,80	0,83	0,82	0.84	0.62
Moy.	0,84	0,55	0,22	0,85	0,93	0,55	0,95	0,63	0,96	0,88	0,95	0,72

Tab. 6.14 – Résultats obtenus par AROMA et les méthodes évaluées lors de la campagne OAEI 2006

6.5.4 Comparaison

Afin de comparer les performances d'AROMA à celles obtenues par les méthodes évaluées lors de la campagne OAEI 2006, nous avons sélectionné Ipee qui obtient globalement les meilleurs résultats et nous avons utilisé la méthode complète avec réduction de la cardinalité. Nous avons choisi un seuil de sélection des règles φ_r égal à $0,5$ (valeur au-dessous de laquelle une règle n'est pas statistiquement significative en terme d'écart à l'équilibre). Les résultats sont présentés par la table 6.14. Pour chaque méthode, les résultats (en terme de précision et rappel) obtenus sur chacune des séries de tests (1xx, 2xx, 3xx) et l'ensemble des tests sont données. Les résultats représentent les moyennes harmoniques des valeurs de précision et de rappel obtenues sur chaque test individuel.

Sur la moyenne des résultats, AROMA est très significativement surpassée par les méthodes Coma, Falcon et RiMOM. Sur la première série, les résultats d'AROMA sont, comme pour toutes les méthodes, très bons. Sur la deuxième série de tests, elle obtient une très bonne précision mais un rappel moyen plus faible. Cependant, seules Coma et RiMOM obtiennent des meilleurs résultats sur ces deux points. La méthode Falcon obtient une moins bonne précision mais un rappel largement plus élevé. Sur la dernière série de tests (3xx), ses résultats sont, tout comme la méthode Coma, très significativement en baisse. En effet, seules les méthodes OCM et H-match obtiennent des résultats moins bons à la fois en terme de précision et de rappel.

Conclusion

Dans ce chapitre, nous avons tout d'abord présenté les réalisations logicielles de notre thèse. Plus particulièrement, nous avons décrit un outil original d'aide à l'alignement qui permet, à un expert, de l'aider dans sa démarche de validation d'un alignement. Etant donné la taille, souvent importante des hiérarchies, la visualisation d'un alignement devient rapidement problématique. Ainsi, nous avons également proposé une série de filtres permettant d'alléger la visualisation afin que l'utilisateur puisse se concentrer sur certaines parties de la hiérarchie

ou certains types de relations de correspondance.

Dans un second temps, nous avons décrit les expérimentations réalisées dans le but d'étudier le comportement et la performance d'AROMA sur différents jeux de tests et avec différentes mesures d'intérêt. Dans un premier temps, nous avons réalisé l'étude de l'approche pré-traitement permettant la sélection et l'association des termes représentatifs. Les résultats montrent, dans ce cas, que la confiance et l'indice de Loevinger (qui sont des mesures asymétriques de nature descriptive) ne sont pas adaptées. Dans ce cas, l'utilisation d'indices probabilistes tels que l'intensité d'implication ou Ipee semble plus adéquate.

La suite des expérimentations a concerné les autres phases (extraction de règles, post-traitements et alignement syntaxique) d'AROMA. Sur l'ensemble des tests, l'utilisation du filtre de réduction de la cardinalité permet d'augmenter très significativement les scores obtenus[8]. L'alignement syntaxique a également un effet très positif. Seul l'alignement des hiérarchies textuelles a tiré bénéfice du filtre d'élimination des inconsistances. Du coté des mesures d'intérêt, l'intensité d'implication obtient les meilleurs résultats dans le cas des hiérarchies textuelles. Pour l'alignement du jeu d'ontologies OWL, c'est Ipee qui se comporte le mieux. D'une manière générale, les écarts entre les scores obtenus par les différentes mesures se resserrent lorsque l'on utilise la méthode complète.

[8]Cette amélioration est en grande partie due au fait que les alignements de références utilisés sont fonctionnels

Conclusion

Les travaux menés dans cette thèse s'insèrent à l'intersection de deux domaines de recherche que sont l'ingénierie des connaissances et l'extraction des connaissances dans les données. Notre objectif a été de tirer profit des travaux menés en fouille de règles d'association, et notamment sur les mesures d'intérêt, dans le but d'aligner des ontologies et également toutes sortes de hiérarchies textuelles (catalogues et répertoires Web, thésaurus, etc).

Le résultat de notre travail est une méthode originale d'alignement dénommée AROMA (Association Rule Matching Approach). L'originalité de notre méthode réside dans la combinaison des caractéristiques suivantes :

- **extensionnelle** : la découverte de l'alignement s'appuie principalement sur le contenu indexé aux hiérarchies (instances, textes, etc...) ;
- **terminologique** : la comparaison de deux entités se fait à partir d'une sélection de termes extraits du contenu textuel ;
- **implicative** : de par la nature asymétrique des règles d'association, notre méthode permet de découvrir non seulement des entités en relation d'équivalence mais également en relation d'implication (spécialisation ou composition).

Son caractère extensionnel et terminologique lui permet d'être relativement générique de par sa faible dépendance vis-à-vis de la sémantique du langage de représentation utilisé. En effet, les seules contraintes d'AROMA concernent l'organisation hiérarchique (par une relation d'ordre partiel) des schémas à aligner et la présence d'un corpus textuel associé aux entités des schémas.

Son caractère implicatif lui permet de produire des alignements plus riches en terme de sémantique que les méthodes d'alignement basées seulement sur des mesures de similarités.

Contributions de la thèse

La méthode d'alignement AROMA constitue la contribution principale de notre thèse. Cette contribution a été accompagnée de propositions connexes tant sur le plan théorique que sur le plan expérimental. Du point de vue théorique, nous avons également proposé :

- une formalisation implicative de l'alignement prenant en compte les aspects de redondance, de symétrie et de cardinalité ;
- l'adaptation d'un modèle d'évaluation permettant de mieux prendre en

compte les alignements implicatifs.

Du point de vue expérimental, nous avons réalisé :

- l'implémentation de la méthode AROMA dans un environnement interactif d'aide à la visualisation, à la validation, et à l'édition d'alignements ;
- des études expérimentales portant d'une part sur l'extraction terminologique, et d'autre part sur le comportement et la performance de la méthode.

Un modèle d'alignement implicatif

A partir de la formalisation d'une hiérarchie, qui est la structure de base de nombreuses représentations (telles que les ontologies, catalogues Web, annuaires de sites, etc.), nous avons proposé un modèle implicatif d'alignement. Nous avons introduit des règles permettant de déduire, à partir de l'alignement et des hiérarchies, de nouveaux éléments de correspondance. Grâce à ces règles, nous avons ensuite défini les notions de fermeture et couverture minimale d'un alignement. Ces dernières notions nous ont permis de formaliser la redondance dans un alignement et sa consistance. Finalement, nous avons étudié la symétrie et la cardinalité d'un alignement.

Une adaptation d'un modèle d'évaluation sémantique pour une meilleure prise en compte des implications

Nous avons montré que le modèle d'évaluation classique des méthodes d'alignement, basé sur les mesures de précision et de rappel, n'est pas du tout adapté aux alignements implicatifs. Partant d'un modèle d'évaluation sémantique proposé par J. Euzenat [Euz07] et de la notion de fermeture présentée dans notre modèle d'alignement, nous proposons une adaptation permettant l'utilisation des mesures de précision et de rappel idéales. Ces mesures ont l'avantage de prendre en compte les capacités déductives de notre modèle d'alignement et permettent ainsi d'évaluer de la même manière deux alignements sémantiquement égaux.

La réalisation logicielle de la méthode AROMA

Les réalisations logicielles de notre thèse ont porté sur le développement de la méthode AROMA et également sur un outil d'aide à la validation des alignements produits. Cet outil vise à accompagner l'utilisateur dans cette démarche de validation en lui proposant une représentation de l'alignement sous forme de graphe, mais également une série de filtres lui permettant d'alléger la représentation visuelle et de focaliser sur certains types de relations ou certaines parties du graphe. Cet outil permet également de visualiser les informations relatives à un élément de correspondance, et de le valider.

Etudes expérimentales

Nous nous sommes également attachés à l'étude expérimentale de notre méthode AROMA sur différents types de hiérarchies et avec différentes mesures

de qualité. Nous avons premièrement étudié la phase de sélection des termes, issue du prétraitement. Cette première étude montre que parmi les mesures asymétriques testées, seules les mesures statistiques ont, dans ce cas, un pouvoir filtrant intéressant. Dans un second temps, nous avons procédé à l'évaluation de 5 versions différentes d'AROMA sur un alignement de hiérarchies textuelles et également sur une série d'alignements d'ontologies (proposée par l'INRIA). Pour chacun de ces tests, nous avons évalué le comportement et la performance en fonction des valeurs de seuil et de la mesure d'intérêt utilisée. Parmi les 6 mesures d'intérêt sélectionnées, l'intensité d'implication donne les meilleurs résultats sur les hiérarchies textuelles. Dans le cas du jeu de tests de l'INRIA, Ipee, la confiance et l'indice de Loevinger sont les mesures les plus adaptées.

Perspectives

Les perspectives de cette thèse concernent en premier lieu l'amélioration de la méthode AROMA tout en conservant l'approche d'alignement implicatif. Nous envisageons deux voies possibles. La première concerne la découverte d'implications au niveau intensionnel. La seconde consiste à étudier la découverte de correspondances complexes du type $a \wedge b \Rightarrow c$.

Ensuite, nous envisageons, à l'instar des applications d'ECD classiques, l'élaboration d'un système anthropocentré d'aide à l'alignement.

Finalement, il pourrait être également intéressant de définir un indice global d'implication entre hiérarchies (ou ontologies).

Découverte d'implications au niveau intensionnel

Pour l'instant, AROMA évalue des implications au niveau extensionnel. Il pourrait être intéressant d'étudier également l'apport des mesures d'intérêt asymétriques au niveau intensionnel. Notamment, on pourrait envisager l'utilisation de ces mesures par des techniques structurelles ou encore terminologiques linguistiques.

Au niveau structurel, on pourrait, à partir de l'alignement des propriétés, utiliser des mesures d'intérêt asymétriques afin de vérifier la tendance inclusive entre les ensembles de propriétés possédées par les concepts. Dans ce cas, la découverte d'implications entre deux concepts suivrait la stratégie suivante : « un concept x est plus général qu'un concept y si l'ensemble des propriétés de x a tendance à être inclus dans l'ensemble des propriétés de y ».

Au niveau terminologique linguistique, le recours à une base de données lexicales (du type Wordnet) et aux mesures d'intérêt pourrait également permettre la découverte d'implications (de manière non-stricte). En effet, nous avons vu dans la section 3.4.1 que les mesures de similarité sémantiques (type Lin, Rada, etc.) sont issues de la composition d'un estimateur de quantité d'information taxonomique et d'une mesure de similarité (type Dice, Jaccard, etc.). Afin de vérifier si un terme a tendance à être plus spécifique qu'un autre, on pourrait combiner une mesure de quasi-implication et un des estimateurs présentés dans la section 3.4.1.

Découverte de correspondances complexes

Classiquement, le modèle des règles d'association est conçu pour identifier des tendances implicatives entre des conjonctions d'attributs. Dans le cadre de notre thèse, nous avons restreint ce modèle à des règles binaires (c.-à-d. n'ayant qu'un seul attribut en prémisse et en conclusion).

Dans le domaine de l'alignement de schémas de bases de données, la prise en compte de correspondances complexes entre attributs a été étudiée à plusieurs reprises [DH05]. Une correspondance complexe est une expression de la forme $a_1 op_1 a2 op_2 ... a_n = b_1 op'_1 ... b_m$ où les a_i et b_j sont des attributs et les op_x sont des opérations (opération arithmétique, ensembliste, logique, concaténation, etc.). Un exemple de telle correspondance est $Prix\ HT \times (1 + Taux\ TVA) = Prix\ TTC$.

Dans le domaine de l'alignement d'ontologies (ou de hiérarchies textuelles), les correspondances complexes ont été très peu étudiées. Un des seuls travaux allant dans ce sens a été celui de A. Doan [DMD+03], qui s'est restreint d'une part, aux disjonctions de concepts, et d'autre part, à la relation d'équivalence.

En partant du principe que certaines entités d'une hiérarchie ont une intersection de leurs extensions non vide, il pourrait être intéressant d'utiliser le modèle des règles d'association pour découvrir des correspondances (implicatives et conjonctives) du type $a \wedge b \Rightarrow c$ (ou $a \Rightarrow c \wedge d$). Les correspondances complexes seraient particulièrement utiles pour aligner une hiérarchie (ou ontologie) supportant l'héritage multiple avec une autre basée uniquement sur de l'héritage simple. Par exemple, on pourrait trouver, dans ce cas, une relation du genre *véhicule terrestre* $\wedge$ *véhicule maritime* $\Rightarrow$ *aéroglisseur*.

Elaboration d'une démarche anthropocentrée d'aide à l'alignement

Dans une démarche d'ECD, l'utilisateur est fortement impliqué dans le processus, tant pour effectuer des choix (pré-traitements à utiliser, paramètres des algorithmes de fouille, etc.), que pour étudier et valider les connaissances produites.

Suivant ce principe, il serait intéressant de développer une démarche complète, visuelle et interactive, d'aide à l'alignement d'ontologies. Afin d'atteindre ce but, on devrait tout d'abord s'intéresser à l'implication de l'utilisateur dans le choix des pré-traitements et des méthodes d'alignement à utiliser. Quelques approches flexibles de ce genre ont été proposées dans le cadre de l'alignement d'ontologies et de schémas (COMA [DR02] et son extension COMA++ [ADMR05]). L'autre aspect qui devrait être davantage étudié est celui de l'explication des alignements produits. En effet, des efforts doivent être faits pour assurer une explication intelligible des résultats pour l'utilisateur afin qu'il puisse mieux comprendre pourquoi une relation a été détectée. Enfin, un dernier aspect concerne la visualisation des alignements qui est rendue très souvent problématique de par la taille des structures. Pour cela, deux approches sont envisageables : (1) l'utilisation de filtres supplémentaires permettant d'alléger la représentation graphique de l'alignement, et (2) le développement d'algorithmes de tracé de graphes adaptés.

Vers un indice global d'implication entre hiérarchies

Enfin, la mise au point d'un indice global d'implication entre hiérarchies (ontologies ou autres représentations) semble être également une voie intéressante. En effet, de par la prolifération des ontologies disponibles, un utilisateur peut être amené, lorsqu'il décide par exemple de faire évoluer une application, à rechercher une ontologie plus précise, ou ayant un degré de généralité supérieur à celle dont il dispose. Dans ce but, on pourrait à partir d'alignements calculés, identifier l'inclusion partielle entre ontologies. Pour ce faire, il serait possible de développer un modèle et des adaptations de mesures d'intérêt visant à quantifier le degré d'inclusion entre ontologies.

Bibliographie

[ADMR05] D. Aumueller, H.-H. Do, S. Massmann et E. Rahm – « Schema and ontology matching with coma++ », *Proceedings of the 2005 ACM SIGMOD international conference on Management of data (SIGMOD 05)* (New York, NY, USA), ACM Press, 2005, p. 906–908.

[Aim07] X. Aimé – « Visualisation interactive d'alignements implicatifs entre hiérarchies conceptuelles », Tech. report, Mémoire d'ingénieur - CNAM Nantes, 2007.

[AIS93] R. Agrawal, T. Imielinski et A. Swami – « Mining association rules between sets of items in large databases », *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ACM Press, 1993, p. 207–216.

[AS94] R. Agrawal et R. Srikant – « Fast algorithms for mining association rules in large databases », *Proceedings of 20th International Conference on Very Large Data Bases (VLDB 94)* (J. B. Bocca, M. Jarke et C. Zaniolo, éds.), Morgan Kaufmann, 1994, p. 487–499.

[AS01] R. Agrawal et R. Srikant – « On integrating catalogs », *Proceedings of the 10th international conference on World Wide Web (WWW 01)* (New York, NY, USA), ACM Press, 2001, p. 603–612.

[AY01] C. C. Aggarwal et P. S. Yu – « Mining associations with the collective strength approach », *IEEE Transactions on Knowledge and Data Engineering* **13** (2001), no. 6, p. 863–873.

[Aze03] J. Aze – « Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances », *Revue des Sciences et Technologies de l'Information* **17** (2003), no. 1-3, p. 171–182, Actes des journées Extraction et Gestion des Connaissances (EGC) 2003.

[BA99] R. J. Bayardo et R. Agrawal – « Mining the most interesting rules », *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining*, ACM Press, 1999, p. 145–154.

[Bac06] T.-L. Bach – « Construction d'un web sémantique multi-points de vue », Thèse, INRIA Sophia Antipolis, 2006.

[BDKG04] T.-L. Bach, R. Dieng-Kuntz et F. Gandon – « On ontology matching problems (for building a corporate semantic web in a multi-communities organization) », *Proceedings of the 6th International Conference on Enterprise Information Systems (ICEIS 2004)* (Porto (PT)), 2004, p. 236–243.

[BHK07] E. Blanchard, M. Harzallah et P. Kuntz – « Vers une classification des similarités basées sur le contenu informationnel des concepts d'une hiérarchie de subsomption », *Actes des 18èmes journées francophones sur l'Ingénierie des Connaissances (IC 2007)*, Cépadues, 2007, p. 145–156.

[BHL01] T. Berners Lee, J. Hendler et O. Lassila – « The Semantic Web », *Scientific American* **284** (2001), no. 5, p. 34–43.

[BKHB06] E. Blanchard, P. Kuntz, M. Harzallah et H. Briand – « A tree-based similarity for evaluating concept proximities in an ontology », *Proceedings of the 10th conference of the International Federation of Classification Societies (ICFCS 2006)*, Springer, 2006, p. 3–11.

[BL04] R. J. Brachman et H. J. Levesque – *Knowledge representation and reasoning*, Elsevier, 2004.

[Bla05] J. Blanchard – « Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association », Thèse, Université de Nantes, 2005.

[BLN86] C. Batini, M. Lenzerini et S. Navathe – « A comparative analysis of methodologies for database schema integration », *ACM Computing Surveys* **18** (1986), no. 4, p. 323–364.

[BMS97] S. Brin, R. Motwani et C. Silverstein – « Beyond market baskets : generalizing association rules to correlations », *SIGMOD Record* **26** (1997), no. 2, p. 265–276.

[BMUT97] S. Brin, R. Motwani, J. D. Ullman et S. Tsur – « Dynamic itemset counting and implication rules for market basket data », *SIGMOD Record* **26** (1997), no. 2, p. 255–264.

[Bou94] D. Bourigault – « Lexter : un logiciel d'extraction de terminologie. application à l'acquisition des connaissances à partir des textes. », Thèse, Ecole des Hautes Etudes en Sciences sociales, Paris, 1994.

[BSGG04] H. Briand, M. Sebag, R. Gras et F. Guillet (éds.) – *Mesures de qualité pour la fouille de données*, Cépaduès Editions, 2004, numéro spécial de la Revue des Nouvelles Technologies de l'Information.

[CAdV01] S. Castano, V. D. Antonellis et S. D. C. di Vimercati – « Global viewing of heterogeneous data sources », *IEEE Transactions on Knowledge and Data Engineering* **13** (2001), no. 2, p. 277–297.

[CFM05] S. Castano, A. Ferrara et S. Montanelli – « Matching ontologies in open networked systems : Techniques and applications », *Journal on Data Semantics (JoDS)* **5** (2005), p. 25–63.

[Che04] H. Cherfi – « Etude et réalisation d'un système d'extraction de connaissances à partir de textes », Thèse, UHP, Nancy 1, Novembre 2004.

[Chu88] K. W. Church – « A stochastic parts program and noun phrase parser for unrestricted text », *Proceedings of the 2nd Conference on Applied Natural Language Processing*, ACL, 1988, p. 136–143.

[CK96] F. Caillez et P. Kuntz – « A contribution to the study of the metric and euclidiean structures of dissimilarities », *Psychometrika* **61** (1996), no. 2, p. 241–253.

[Coh60] J. Cohen – « A coefficient of agreement for nominal scales », *Educational and Psychological Measurement* **20** (1960), no. 1, p. 37–46.

[CR06] A. Ceglar et J. F. Roddick – « Association mining », *ACM Computing Surveys* **38** (2006), no. 2, p. 5.

[CRF03] W. Cohen, P. Ravikumar et S. Fienberg – « A comparison of string metrics for matching names and records », *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb 03)*, 2003, p. 73–78.

[Dai94] B. Daille – « Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques », Thèse, Université Paris 7, 1994.

[Dai03] — , « Conceptual structuring through term variations », *Proceedings ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment* (F. Bond, A. Korhonen, D. MacCarthy et A. Villacicencio, éds.), 2003, p. 9–16.

[DGB06] J. David, F. Guillet et H. Briand – « Matching directories and owl ontologies with aroma », *Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM 06)* (New York, NY, USA), ACM Press, 2006, p. 830–831.

[DGB07a] J. David, F. Guillet et H. Briand – « Association rule ontology matching approach », *International Journal on Semantic Web and Information Systems* **3** (2007), no. 2, p. 27–49.

[DGB07b] — , « Comparaison de mesures d'intérêt pour l'alignement de hiérarchies textuelles », *Actes des 18ème journées francophones sur l'ingénierie des connaissances (IC 2007)* (F. trichet, éd.), Cépaduès, Juillet 2007, p. 13–24.

[DGGB06a] J. David, F. Guillet, R. Gras et H. Briand – « Alignement de hiérarchies conceptuelles par découverte d'implications entre concepts », *Revue des Nouvelles Technologies de l'Information* **E-1** (2006), p. 151–162, Actes de la Conférence EGC'06.

[DGGB06b] — , « Conceptual hierarchies matching : an approach based on the discovery of implication rules between concepts », *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI 06)* (Riva del Garda, Italy) (G. Brewka, S. Coradeschi, A. Perini et P. Traverso, éds.), IOS Press, august 2006, p. 357–361.

[DGGB07] — , « Comparison of interestingness measures applied to textual taxonomies matching », *Proceedings of the 12th symposium on Applied Stochastic Models and Data Analysis (ASMDA 07)*, 2007, p. To appear.

[DGP+05a] J. David, F. Guillet, V. Philippé, H. Briand et R. Gras – « Validation d'une expertise textuelle basée sur l'intensité d'implication », *Atelier DKQ de la conférence Extraction et Gestion des Connaissances 2005*, 2005.

[DGP+05b] J. David, F. Guillet, V. Philippé, R. Gras et H. Briand – « Validation d'une expertise textuelle par une méthode de classification basée sur l'intensité d'implication », *Actes des Rencontres Analyse Statistique Implicative* (F. Spagnolo, R. Gras et J. David, éds.), 2005, p. 157–162.

[DGPG05] J. David, F. Guillet, V. Philippé et R. Gras – « Implicative statistical analysis applied to clustering of terms taken from a psychological text corpus », *Proceedings of the 11th symposium on Applied Stochastic Models and Data Analysis (ASMDA 05)* (Brest, France) (J. Janssen et P. Lenca, éds.), 2005, p. 201–208.

[DH05] A. Doan et A. Y. Halevy – « Semantic-integration research in the database community », *AI Magazine* **26** (2005), no. 1, p. 83–94.

[Dic45] L. Dice – « Measures of the amount of ecologic association between species », *Ecology* **26** (1945), no. 3, p. 297–302.

[DMD+03] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos et A. Halevy – « Learning to match ontologies on the semantic web », *The VLDB Journal* **12** (2003), no. 4, p. 303–319.

[DMDH02] A. Doan, J. Madhavan, P. Domingos et A. Halevy – « Learning to map between ontologies on the semantic web », *Proceedings of the 11th international conference on World Wide Web (WWW 02)* (New York, NY, USA), ACM Press, 2002, p. 662–673.

[DMDH04] A. Doan, J. Madhavan, P. Domingos et A. Halevy – « Ontology matching : a machine learning approach », Handbook on Ontologies in Information Systems (S. Staab et R. Studer, éds.), Springer-Velag, 2004, p. 397–416.

[DR02] H. Do et E. Rahm – « Coma - a system for flexible combination of schema matching approaches », *Proceedings of the 28th International Conference on Very Large Data Bases (VLDB 02)*, 2002, p. 610–621.

[EBB+04] J. Euzenat, T. L. Bach, J. Barrasa, P. Bouquet, J. D. Bo, R. Dieng-Kuntz, M. Ehrig, M. Hauswirth, M. Jarrar, R. Lara, D. Maynard, A. Napoli, G. Stamou, H. Stuckenschmidt, P. Shvaiko, S. Tessaris, S. V. Acker et I. Zaihrayeu – « State of the art on ontology alignment », Deliverable D2.2.3, Knowledge Web NoE, 2004.

[EE05] M. Ehrig et J. Euzenat – « Relaxed precision and recall for ontology matching », *Proceedings of the Workshop on Integrating Ontologies* (B. Ashpole, M. Ehrig, J. Euzenat et H. Stuckenschmidt, éds.), vol. 156, CEUR-WS.org, 2005, p. 25–32.

[ES04] M. Ehrig et S. Staab – « QOM – quick ontology mapping », *Proceedings of the 3rd International Semantic Web Conference (ISWC 2004)* (Hiroshima (JP)), Lecture notes in computer science, vol. 3298, 2004, p. 683–697.

[ES07] J. Euzenat et P. Shvaiko – *Ontology matching*, Springer-Verlag, Heidelberg (DE), 2007 (english).

[Euz07] J. Euzenat – « Semantic precision and recall for ontology alignment evaluation », *Proceedings of 20th International Joint Confe-*

rence on Artificial Intelligence (IJCAI 07) (Hyderabad (IN)), 2007, p. 248–253.

[EV04] J. Euzenat et P. Valtchev – « Similarity-based ontology alignment in owl-lite », *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 04)*, 2004, p. 333–337.

[FPSM91] W. J. Frawley, G. Piatetsky-Shapiro et C. J. Matheus – « Knowledge discovery in databases : An overview », Knowledge Discovery in Databases, AAAI/MIT Press, 1991, p. 1–30.

[FPSS96] U. M. Fayyad, G. Piatetsky-Shapiro et P. Smyth – « From data mining to knowledge discovery : An overview », Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, 1996, p. 1–34.

[Fre98] A. A. Freitas – « On objective measures of rule surprisingness », *Proceedings of the 2nd European conference on principles of data mining and knowledge discovery (PKDD 98)* (J. Zytkow et M. Quafafou, éds.), Lecture Notes in Artificial Intelligence, vol. 1510, Springer-Verlag, 1998, p. 1–9.

[Gan91] J.-G. Ganascia – « Charade : apprentissage de bases de connaissances », Induction symbolique et numérique à partir de données (Y. Kodratoff et E. Diday, éds.), Cépaduès Editions, 1991, p. 309–326.

[GCB+04] R. Gras, R. Couturier, J. Blanchard, H. Briand, P. Kuntz et P. Peter – « Quelques critères pour une mesure de qualité de règles d'association », *Revue des Nouvelles Technologies de l'Information* **E-1** (2004), p. 3–31, numéro spécial Mesures de qualité pour la fouille de données.

[GDRG06] R. Gras, J. David, J. Régnier et F. Guillet – « Typicalité et contribution des sujets et des variables supplémentaires en analyse statistique implicative », *Revue des Nouvelles Technologies de l'Information* **E-1** (2006), p. 151–162, Actes de la Conférence EGC 06.

[GF98] D. A. Grossman et O. Frieder – *Information retrieval : Algorithms and heuristics*, Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[GH07] F. Guillet et H. J. Hamilton – *Quality measures in data mining (studies in computational intelligence)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[Gra96] R. Gras et al. – *L'implication statistique, une nouvelle méthode exploratoire de données*, La pensée sauvage, 1996.

[Gru93] T. Gruber – « A translation approach to portable ontologies », *Knowledge Acquisition* **5** (1993), no. 2, p. 199–220.

[GSY04] F. Giunchiglia, P. Shvaiko et M. Yatskevich – « S-match : an algorithm and an implementation of semantic matching », *Proceedings of European Semantic Web Symposium*, LNCS 3053, 2004, p. 61–75.

[Gui04] F. Guillet – *Mesures de la qualité des connaissances en ECD*, 2004, Tutoriel des journées Extraction et Gestion des

Connaissances (EGC 04), www.isima.fr/~egc2004/Cours/Tutoriel-EGC2004.pdf.

[Ham50] R. Hamming – « Error-detecting and error-correcting codes », *The Bell System Technical Journal* **26** (1950), no. 2, p. 147–160.

[HaYQW05] W. Hu, N. J. ans Y.Z. Qu et Y. Wang – « Gmo : A graph matching for ontologies », *K-Cap 2005 Workshop on Integrating Ontologies*, 2005, p. 43–50.

[HF99] J. Han et Y. Fu – « Mining multiple-level association rules in large databases », *IEEE Transactions on Knowledge and Data Engineering* **11** (1999), no. 5, p. 798–805.

[HG04] J. Hayes et C. Gutiérrez – « Bipartite graphs as intermediate model for rdf », *Proceedings of the 3rd International Semantic Web Conference (ISWC 04)*, 2004, p. 47–61.

[HH01] R. Hilderman et H. Hamilton – *Knowledge discovery and measures of interestingness*, Kluwer Academic, 2001.

[HYNT04] T. Hoshiai, Y. Yamane, D. Nakamura et H. Tsuda – « A semantic category matching approach to ontologies alignment », *Proceedings of the 3rd international workshop on Evaluation of Ontology Based Tools (EON 2004)*, 2004.

[IDC07] « The expanding digital universe : A forecast of worldwide information growth through 2010. » – IDC white paper - Sponsored by EMC, Mars 2007.

[IHT04] R. Ichise, M. Hamasaki et H. Takeda – « Discovering relationships among catalogs », *Proceedings of the 7th International Conference on Discovery Science (DS 04)* (E. Suzuki et S. Arikawa, éds.), LNCS, vol. 3245, Springer, 2004, p. 371–379.

[Jac01] P. Jaccard – « Etude comparative de la distribution florale dans une portion des Alpes et du Jura », *Bulletin de la Société Vaudoise des Sciences Naturelles* **37** (1901), p. 547–579.

[Jar89] M. A. Jaro – « Advances in record linking methodology as applied to the 1985 census of tampa florida », *Journal of the American Statistical Association* **84** (1989), no. 406, p. 414–420.

[KS03] Y. Kalfoglou et M. Schorlemmer – « Ontology mapping : the state of the art », *Knowledge Engineering Review* **18** (2003), no. 1, p. 1–31.

[Kul27] S. Kulczynski – « Die pflanzenassoziationen der pieninen », *Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des Sciences Mathématiques et Naturelles* **B** (1927), no. suppl. 2, p. 57–203.

[KVS06] K. Kotis, G. Vouros et K. Stergiou – « Towards automatic merging of domain ontologies : The HCONE-merge approach », *Journal of Web Semantics* **4** (2006), no. 1, p. 60–79.

[Leh00] R. Lehn – « Un système interactif de visualisation et de fouille de règles pour l'extraction de connaissances dans les bases de données », Thèse de doctorat, Université de Nantes, 2000.

[Ler81] I.-C. Lerman – *Classification et analyse ordinale des données*, Dunod, 1981.

[Lev66] V. I. Levenshtein – « Binary codes capable of correcting deletions, insertions, and reversals », *Soviet Physics Doklady* **10** (1966), p. 707–710.

[LFZ99] N. Lavrac, P. A. Flach et B. Zupan – « Rule evaluation measures : a unifying view », *Proceedings of the ninth International Workshop on Inductive Logic Programming (ILP 99)*, Springer-Verlag, 1999, p. 174–185.

[LG01] M. S. Lacher et G. Groh – « Facilitating the exchange of explicit knowledge through ontology mappings », *Proceedings of the 14th International Florida Artificial Intelligence Research Society Conference (FLAIRS 01)*, AAAI Press, 2001, p. 305–309.

[Lin98] D. Lin – « An information-theoretic definition of similarity », *Proceedings of the 15th international conference on machine learning*, Morgan Kaufmann, 1998, p. 296–304.

[Loe47] J. Loevinger – « A systematic approach to the construction and evaluation of tests of ability », *Psychological Monographs* **61** (1947), no. 4.

[LS88] L. Lebart et A. Salem – *Analyse statistique des données textuelles. questions ouvertes et lexicométrie*, Dunod, 1988.

[LT04] S. Lallich et O. Teytaud – « Evaluation et validation de l'intérêt des règles d'association », *Revue des Nouvelles Technologies de l'Information* **E-1** (2004), p. 193–218, numéro spécial Mesures de qualité pour la fouille de données.

[MBR01] J. Madhavan, P. A. Bernstein et E. Rahm – « Generic schema matching with cupid », *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB 01)*, 2001, p. 49–58.

[McB02] B. McBride – « Jena : A semantic web toolkit », *IEEE Internet Computing* **06** (2002), no. 6, p. 55–59.

[MCP01] G. Minnen, J. Carroll et D. Pearce – « Applied morphological processing of english », *Natural Language Engineering* **7** (2001), no. 3, p. 207–223.

[MGMR02] S. Melnik, H. Garcia-Molina et E. Rahm – « Similarity flooding : A versatile graph matching algorithm and its application to schema matching », *Proceedings of the 18th International Conference on Data Engineering (ICDE 02)*, IEEE Computer Society, 2002, p. 117–128.

[Mit97] T. Mitchell – *Machine learning*, McGraw-Hill, 1997.

[MMS93] M. P. Marcus, M. A. Marcinkiewicz et B. Santorini – « Building a large annotated corpus of english : the penn treebank », *Computational Linguistics* **19** (1993), no. 2, p. 313–330.

[Mos68] F. Mosteller – « Association and estimation in contingency tables », *Journal of the American Statistical Association* **63** (1968), no. 321, p. 1–28.

[MZ02] A. Mäedche et V. Zacharias – « Clustering ontology-based metadata in the semantic web », *the proccedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 02)*, 2002, p. 348–360.

[NM01] N. F. Noy et M. A. Musen – « Anchor-prompt : using non-local context for semantic matching », *Workshop on Ontologies and Information Sharing at the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 01)*, 2001.

[NM03] — , « The prompt suite : interactive tools for ontology merging and mapping », *International Journal of Human-Computer Studies* **59** (2003), no. 6, p. 983–1024.

[NS06] H. Nottelmann et U. Straccia – « A probabilistic, logic-based framework for automated web directory alignment », Soft Computing in Ontologies and the Semantic Web (Z. Ma, éd.), Studies in Fuzziness and Soft Computing, Springer Verlag, 2006, p. 47–77.

[NW70] S. B. Needleman et C. D. Wunsch – « A general method applicable to the search for similarities in the amino acid sequence of two proteins », *Journal of Molecular Biology* **48** (1970), no. 3, p. 443–453.

[NWL+04] P. Naïm, P.-H. Wuillemin, P. Leray, O. Pourret et A. Becker – *Réseaux bayésiens*, Eyrolles, Paris, 2004.

[Och57] A. Ochiai – « Zoogeographic studies on the soleoid fishes found in japan ans its neighbouring regions », *Bulletin of the Japanese Society of Scientific Fisheries* **22** (1957), p. 526–530.

[Pea96] K. Pearson – « Mathematical contributions to the theory of evolution : regression, heredity and panmixia », *Philosophical Transactions of the Royal Society Of London* **series A** (1896), no. 187, p. 253–318.

[PL02] P. Pantel et D. Lin – « Discovering word senses from text », *Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 02)* (New York, NY, USA), ACM Press, 2002, p. 613–619.

[PS91] G. Piatetsky-Shapiro – « Discovery, analysis, and presentation of strong rules », Knowledge Discovery in Databases (G. Piatetsky-Shapiro et W. J. Frawley, éds.), AAAI/MIT Press, 1991, p. 229–248.

[PSU98] L. Palopoli, D. Saccá et D. Ursino – « An automatic technique for detecting type conflicts in database schemes », *Proceedings of the 17th international conference on Information and knowledge management (CIKM 98)* (New York, NY, USA), ACM Press, 1998, p. 306–313.

[PTB+05] N. Pasquier, R. Taouil, Y. Bastide, G. Stumme et L. Lakhal – « Generating a condensed representation for association rules », *Journal of Intelligent Information Systems* **24** (2005), no. 1, p. 29–60.

[PTU03] L. Palopoli, G. Terracina et D. Ursino – « Experiences using dike, a system for supporting cooperative information system and

data warehouse design », *Information Systems* **28** (2003), no. 7, p. 835–865.

[QHC06] Y. Qu, W. Hu et G. Chen – « Constructing virtual documents for ontology matching », *Proceedings of the 15th International World Wide Web Conference (WWW 06)* (Edinburgh (UK)), 2006, p. 23–31.

[RB01] E. Rahm et P. A. Bernstein – « A survey of approaches to automatic schema matching », *The VLDB Journal* **10** (2001), no. 4, p. 334–350.

[Res95] P. Resnik – « Using information content to evaluate semantic similarity in a taxonomy », *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 95)*, vol. 1, 1995, p. 448–453.

[RMBB89] R. Rada, H. Mili, E. Bicknell et M. Blettner – « Development and application of a metric on semantic nets », *IEEE Transactions on Systems, Man, and Cybernetics* **1** (1989), no. 19, p. 17–30.

[RR40] P. Russel et T. Rao – « On habitat and association of species of anopheline larvae in south-eastern madras », *Journal of the Malaria Institute of India* **3** (1940), p. 153–178.

[RSK06] C. Reynaud, B. Safar et H. Kefi – « Structural techniques for alignment of structurally dissymmetric taxonomies », *Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW 06)* (H. S. Pinto et M. Labsky, éds.), 2006, p. 39–40.

[RT60] D. Rogers et T. Tanimoto – « A computer program for classifying plants », *Science* **132** (1960), no. 3434, p. 1115–1118.

[SA95] R. Srikant et R. Agrawal – « Mining generalized association rules », *Proceedings of the 21st International Conference on Very Large Databases (VLDB 95)*, 1995, p. 407–419.

[SA97] R. Srikant et R. Agrawal – « Mining generalized association rules », *Future Generation Computer Systems* **13** (1997), no. 2-3, p. 161–180.

[SE05] P. Shvaiko et J. Euzenat – « A survey of schema-based matching approaches », *Journal on Data Semantics* **4** (2005), no. LNCS 3730, p. 146–171.

[SEN+06] P. Shvaiko, J. Euzenat, N. Noy, H. Stuckenschmidt, R. Benjamins et M. Uschold (éds.) – *Ontology matching proceedings of the iswc'06 international workshop on ontology matching*, 2006.

[Sha48] C. E. Shannon – « A mathematical theory of communication », *Bell System Technical Journal* **27** (1948), p. 379–423.

[SM58] R. Sokal et C. Michener – « A statistical method for evaluating systematic relationships », *University of Kansas Science Bulletin* **38** (1958), p. 1409–1438.

[Sma93] F. Smadja – « Retrieving collocations from text : Xtract », *Comput. Linguist.* **19** (1993), no. 1, p. 143–177.

[SMS+01] N. Stojanovic, A. Maedche, S. Staab, R. Studer et Y. Sure – « Seal : a framework for developing semantic portals », *Proceedings of the 1st international conference on Knowledge capture (K-CAP 01)* (New York, NY, USA), ACM Press, 2001, p. 155–162.

[SS88] M. Sebag et M. Schoenauer – « Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases », *Proceedings of the European knowledge acquisition workshop (EKAW 88)*, Gesellschaft fÃ¼r Mathematik und Datenverarbeitung mbH, 1988, p. 28.1–28.20.

[ST05] U. Straccia et R. Troncy – « oMAP : Combining classifiers for aligning automatically OWL ontologies », *Proceedings of the 6th International Conference on Web Information Systems Engineering (WISE 2005)* (New York (NY US)), 2005, p. 133–147.

[SW81] T. F. Smith et M. S. Waterman – « Identification of common molecular subsequences », *Journal of Molecular Biology* **147** (1981), no. 1, p. 195–197.

[SWY75] G. Salton, A. Wong et C. S. Yang – « A vector space model for automatic indexing », *Communication of the ACM* **18** (1975), no. 11, p. 613–620.

[TKMS03] K. Toutanova, D. Klein, C. D. Manning et Y. Singer – « Feature-rich part-of-speech tagging with a cyclic dependency network », *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 03)* (Morristown, NJ, USA), Association for Computational Linguistics, 2003, p. 173–180.

[TKS04] P. Tan, V. Kumar et J. Srivastava – « Selecting the right objective measure for association analysis », *Information Systems* **29** (2004), no. 4, p. 293–313.

[TLL+06] J. Tang, J. Li, B. Liang, X. Huang, Y. Li et K. Wang – « Using bayesian decision for ontology mapping », *Journal of Web Semantics* **4** (2006), no. 1, p. 243–262.

[TM00] K. Toutanova et C. D. Manning – « Enriching the knowledge sources used in a maximum entropy part-of-speech tagger », *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora (EMNLP/VLC 2000)* (Morristown, NJ, USA), Association for Computational Linguistics, 2000, p. 63–70.

[Val99] P. Valtchev – « Construction automatique de taxonomies pour l'aide à la représentation de connaissances par objets », Thèse, Université Joseph Fourier - Grenoble, 1999.

[vR79] C. J. van Rijsbergen – *Information retrieval*, 2 éd., Butterworths, London, 1979.

[W3C88] W3C – « Extensible markup language (xml) 1.0 (w3c recommendation), http ://www.w3.org/tr/2000/rec-xml-20001006 », Février 1988.

[W3C04a] — , « Owl web ontology language reference, http ://www.w3.org/tr/owl-ref/ », 2004.

[W3C04b] — , « Rdf primer, http ://www.w3.org/tr/rdf-primer/ », 2004.

[Win99] W. E. Winkler – « The state of record linkage and current research problems », Tech. report, Statistical Research Division, U.S. Bureau of the Census, 1999.

[WP94] Z. Wu et M. Palmer – « Verb semantics and lexical selection », *Proceedings of the 32nd annual meeting of the associations for Computational Linguistics*, 1994, p. 133–138.

[Yul00] G. Yule – « On the association of attributes in statistics », *Philosophical Transactions of the Royal Society of London* **series A** (1900), no. 194, p. 257–319.

[ZR00] D. Zighed et R. Rakotomalala – *Graphes d'induction - apprentissage et data mining*, Hermes, 2000.

Résumé :
Ce travail de thèse s'inscrit à l'intersection des deux domaines de recherche que sont l'extraction des connaissances dans les données (ECD) et de l'ingénierie des connaissances. Plus précisément, en nous appuyant sur la combinaison des travaux menés, d'une part sur l'alignement des ontologies, et d'autre part sur la fouille de règles d'association, nous proposons une nouvelle méthode d'alignement d'ontologies associées à des corpus textuels (taxonomies, hiérarchies documentaires, thésaurus, répertoires ou catalogues Web), appelée AROMA (*Association Rule Matching Approach*).
Dans la littérature, la plupart des travaux traitant des méthodes d'alignement d'ontologies ou de schémas s'appuient sur une définition intentionnelle des schémas et utilisent des relations basées sur des mesures de similarité qui ont la particularité d'être symétriques (équivalences). Afin d'améliorer les méthodes d'alignement, et en nous inspirant des travaux sur la découverte de règles d'association, des mesures de qualité associées, et sur l'analyse statistique implicative, nous proposons de découvrir des appariements asymétriques (implications) entre ontologies. Ainsi, la contribution principale de cette thèse concerne la conception d'une méthode d'alignement extensionnelle et orientée basée sur la découverte des implications significatives entre deux hiérarchies plantées dans un corpus textuel. Notre méthode d'alignement se décompose en trois phases successives. La phase de prétraitement permet de préparer les ontologies à l'alignement en les redéfinissant sur un ensemble commun de termes extraits des textes et sélectionnés statistiquement. La phase de fouille extrait un alignement implicatif entre hiérarchies. La dernière phase de post-traitement des résultats permet de produire des alignements consistants et minimaux (selon un critère de redondance).
Les principaux apports de cette thèse sont : (1) Une modélisation de l'alignement étendue pour la prise en compte de l'implication. Nous définissons les notions de fermeture et couverture d'un alignement permettant de formaliser la redondance et la consistance d'un alignement. Nous étudions également la symétricité et les cardinalités d'un alignement. (2) La réalisation de la méthode AROMA et d'une interface d'aide à la validation d'alignements. (3) Une extension d'un modèle d'évaluation sémantique pour la prise en compte de la présence d'implications dans un alignement. (4) L'étude du comportement et de la performance d'AROMA sur différents types de jeux de tests (annuaires Web, catalogues et ontologies au format OWL) avec une sélection de six mesures de qualité.
Les résultats obtenus sont prometteurs car ils montrent la complémentarité de notre méthode avec les approches existantes.

Mots-clés :
Alignement d'ontologies, Ingénierie des Connaissances, Extraction des Connaissances dans les bases de Données, fouille de données, règle d'association, Web Sémantique, mesures de qualité

Abstract :
This thesis deals with Knowledge Engineering and Knowledge Discovery in Databases (KDD). More precisely, by using the association rule model, we propose a new matching method designed to match ontologies provided with textual data (i.e. thesaurus, web directories, catalogues etc.).
In the literature, most ontology or schema matching approaches rely on similarity measures and, consequently their vast majority is restricted to finding equivalence relations only. In this context, we propose to use the asymmetric nature of the association rule model, of interestingness measures, and of the implicative statistical analysis in order to overcome the restrictions of only-similarity based approaches. The main contribution of this thesis is the introduction of an extensional and asymmetric matching method based on the discovery of significant implication rules between two textual hierarchies.
Our method follows a three-step KDD process : First, the pre-processing step reindexes ontologies on a common set of terms extracted from textual data ; Next, the association rule discovery aims at finding a set of implications between hierarchies ; And finally, the post-processing step allows to provide consistant and minimal (non-redundant) alignments.
The other four contributions of this thesis are : (1) an extended model of alignment dealing with implication. We define the notions of the closure and the minimal cover of an alignment so as formalize its redundancy and consistancy. We also discuss the symmetricity and cardinality of alignements. (2) the implementations of AROMA and AROMAViz supporting the validation of alignements. (3) an extension of a semantic evaluation model taking the implications into account. (4) the study of the efficiency and the behaviour of AROMA obtained on several benchmarks (web directories, catalogues and OWL ontologies) with the use of a selection of six interestingness measures.
The obtained results are promising because they underly the complementarity of our approach with existing ones.

Keywords :
Ontology matching, Knowledge Engineering, Knowledge Discovery in Databases, data-mining, association rule, Semantic Web, interestingness measures